NOT ALL PREFERENCES ARE STRICT: TIE-AWARE PREFERENCE OPTIMIZATION FOR MULTI-IMAGE MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization methods for multimodal large language models (MLLMs) universally assume strict pairwise preferences. We show that this assumption systematically fails in multi-image settings, where responses frequently *tie*—arriving at the same correct conclusion through different image subsets, visual cues, or levels of detail. We formalize a taxonomy of four structurally distinct tie categories—detail, style/framing, evidence-path, and order-invariance ties—among which evidence-path ties are unique to multi-image reasoning and order-invariance ties predominantly arise in multi-image contexts. Forcing such ties into binary labels introduces spurious preference boundaries that manifest as verbosity bias, position bias, and evidence-path overcommitment. We propose **TieDPO**, a tie-aware preference optimization objective that preserves standard DPO on strict pairs while regularizing tie pairs toward indifference via a margin-based penalty. To support training, we construct **MultiTie-22k**, a multi-image preference dataset with ternary labels and structured tie-category annotations. We further introduce **TieBench-MI**, a diagnostic benchmark of 2,400 pairs for evaluating non-strict preference handling. Experiments show that TieDPO matches or exceeds DPO baselines on standard multi-image and single-image tasks while significantly improving tie recognition and reducing preference biases.

1 INTRODUCTION

Preference optimization has become the de facto approach for aligning large language models (LLMs) and multimodal large language models (MLLMs) with human-desired behavior (Rafailov et al., 2023; Ouyang et al., 2022). Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its multimodal extensions (Sun et al., 2024; Li et al., 2023; Yu et al., 2024) optimize a policy directly from pairwise preference data, and all such methods share a foundational assumption: preferences are *strict*—given two candidate responses, one is always chosen and the other rejected.

This assumption is frequently violated in practice, and the violation is *systematic* in multi-image settings. A single-image question admits a narrow range of acceptable answers grounded in one visual input. Multi-image reasoning (Li et al., 2024a; Jiang et al., 2024; Wang et al., 2024a) aggregates information across images, creating a combinatorial space of valid reasoning trajectories: different image subsets, visual cues, and levels of detail can each lead to equally correct and well-grounded conclusions. Forcing two such responses into a binary chosen/rejected label then produces supervision that is *systematically misleading*, teaching the model to prefer one arbitrary evidence path over another equally valid one.

Figure 1 illustrates this: DPO forces a binary label whose gradient accumulates into systematic biases—position overcommitment (Tian et al., 2025), verbosity (Zheng et al., 2023), and fixation on a single reasoning trajectory. Ties are well-documented in pairwise preference evaluation—about one-quarter of Chatbot Arena battles end in ties (Chiang et al., 2024)—and in multi-image reasoning the combinatorial structure of valid evidence paths makes them *structurally inevitable*.

A closer look reveals a deeper problem: binary preference training does not simply fail to learn ties; it degrades the little tie-handling structure the base model has and replaces it with surface preferences. Untuned multi-image MLLMs barely separate tie from strict pairs (Tie-vs-Strict AUC 0.541 / 0.559,

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.

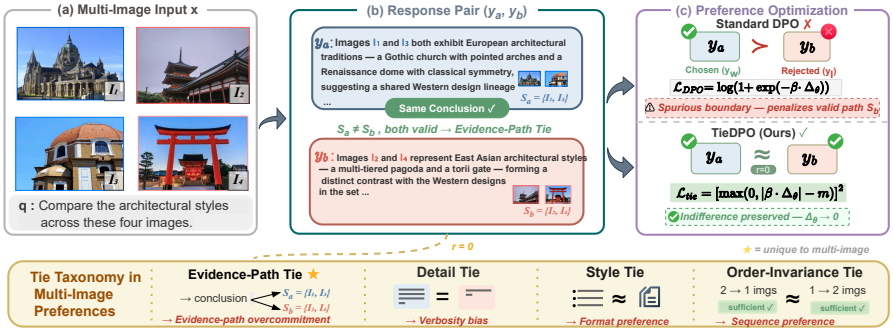


Figure 1: **Ties carry structure that binary labels discard.** Given four architectural images and the query "Compare the architectural styles," Response A draws evidence from $\{I_1, I_3\}$ (European) while Response B draws from $\{I_2, I_4\}$ (East Asian); both reach the same correct conclusion through equally valid evidence paths ($\mathcal{S}_a \neq \mathcal{S}_b$). Standard DPO forces a binary label, creating a spurious preference boundary that penalizes a valid evidence path; TieDPO recognizes this pair as a tie ($r = 0$) and regularizes toward indifference ($s_\theta \to 0$), preserving both paths. The bottom strip illustrates our four-category tie taxonomy, with evidence-path ties (★) unique to multi-image reasoning.

chance 0.5), but the structure they do have is at least calibrated (over-confidence 0.008 / 0.009 on TieBench-MI). After DPO on strict pairs, the separation sharpens (AUC 0.706 / 0.714) but the score distribution becomes badly miscalibrated, with over-confidence rising to 0.077 / 0.066. The newly learned margin is allocated along surface features: on equally good pairs, a strict-only model moves its preference toward a longer response 36.1% of the time against the base model's 27.1%, with image-position and format sensitivities growing in step (Section 5.4). This motivates the research question: *should preference optimization treat ties as a first-class relation whose calibration is a target, rather than noise to discard or binarize?*

Our answer is affirmative. Ties in multi-image data are *structured*: we identify four categories (detail, style/framing, evidence-path, and order-invariance), with **evidence-path ties unique to multi-image reasoning**, and each category induces a specific failure mode when binarized (verbosity, format preference, evidence-path overcommitment, ordering bias; Section 3).

We propose **TieDPO** (**Tie**-aware **D**irect **P**reference **O**ptimization): preserve the DPO objective on strict pairs, and add a margin-based regularizer on tie pairs that penalizes separation in the policy's own preference score. The design target is calibration preservation rather than margin maximization: TieDPO deliberately produces a strict margin that is sufficient to discriminate, rather than the saturated one DPO-strict optimizes for. The tie penalty is reference-free: its fixed point is indifference in the policy itself, so tie pairs are not resolved by whichever response the reference model happens to favor. A margin m defines an indifference zone in which no penalty applies. The method is drop-in, with two extra hyperparameters (λ_{tie}, m); our core contributions are the problem identification and the evaluation framework rather than algorithmic novelty.

To support tie-aware training and evaluation, we construct **MultiTie-22k** (22,198 response pairs with ternary labels; ties are 25.8% of the data, Section 4.3) and **TieBench-MI** (2,400 pairs in two splits, with threshold-free metrics that measure how well a model separates tie from strict pairs and how honestly it scales that separation, Section 4.3).

Our contributions:

- We formalize the **structured tie phenomenon** in multi-image preference data (four categories with distinct origins and binarization failure modes) and document a DPO pathology on this data: strict-only preference training **miscalibrates the tie/strict distinction** and **encodes surface features** (length, image position, discourse form) into the learned preference.
- We propose **TieDPO** (reference-free tie penalty) and release **MultiTie-22k** and **TieBench-MI**; across 7 benchmarks and 2 backbones, TieDPO is the only method that improves tie/strict discrimination without paying for it in calibration, and the only one that reduces all three surface biases below the untuned base.

2 Related Work

2.1 Preference Optimization for LLMs and MLLMs

DPO (Rafailov et al., 2023) reformulates RLHF as a direct policy optimization problem on pairwise preference data, eliminating the explicit reward model; variants including IPO (Azar et al., 2024), SimPO (Meng et al., 2024b), ORPO (Hong et al., 2024), and noise-robust cDPO (Mitchell) modify the objective while retaining the binary preference assumption, as do multimodal extensions (Sun et al., 2024; Li et al., 2023; Yu et al., 2024; Zhao et al., 2023; Ouali et al., 2024; Zhang et al., 2024; Fu et al., 2025) and a visual-side family (POVID (Zhou et al., 2024), V-DPO (Xie et al., 2024), VistaDPO (Huang et al., 2025)). For multi-image alignment, MIA-DPO (Liu et al., 2024c) anchors prompts to individual images and CcDPO (Li et al., 2025) supervises a context-to-cue path; both retain the strict-preference framing and neither addresses the documented position-bias failure mode (Tian et al., 2025). Our work is orthogonal: we extend the preference *relation* itself to include ties, and the tie-aware objective composes with any of these variants.

2.2 Multi-Image MLLMs

Recent multi-image MLLMs (Li et al., 2024a; Liu et al., 2024a; Jiang et al., 2024; Wang et al., 2024b; Chen et al., 2024; Bai et al., 2025; Li et al., 2024b; Ye et al., 2024) demonstrate strong cross-image reasoning, evaluated by a range of multi-image benchmarks (MuirBench (Wang et al., 2024a), MIRB (Zhao et al., 2024), BLINK (Fu et al., 2024), MIBench (Liu et al., 2024b), ReMI (Kazemi et al., 2024), MMIU (Meng et al., 2024a)). A key characteristic of multi-image tasks is that they naturally admit multiple valid reasoning trajectories—different image subsets can support the same conclusion—which is precisely what gives rise to evidence-path ties.

2.3 Ambiguous and Non-Strict Preferences

Ties and non-strict preferences are acknowledged in prior work: classical Bradley–Terry extensions model ties explicitly (Fishburn, 1970; Rao & Kupper, 1967; Davidson, 1970); recent RLHF work explores per-response desirability (Ethayarajh et al., 2024), distributional preferences (Melnyk et al., 2024), and annotator disagreement (Gordon et al., 2022). All of these treat preference ambiguity as a *global, unstructured* property of the supervision signal and apply uniform softening everywhere—including on genuinely strict pairs. We instead treat ties as a *structured, categorical* phenomenon: the taxonomy (Section 3) grounds *when* and *why* regularization applies, leaving strict pairs untouched. Empirically, we further show (Section 5) that standard DPO on multi-image data *degrades base-model preference calibration*, which reframes the goal of tie-aware methods as preserving the calibration structure the base model already provides.

3 Tie Taxonomy in Multi-Image Preference Data

Ties in multi-image preference data are *structured*: we identify four categories (Table 1), each rooted in a different aspect of multi-image reasoning and each inducing a specific failure mode when forcibly binarized. Detail and style/framing ties have analogues in single-image and text-only settings (Zheng et al., 2023); our contribution is to show how they are amplified in multi-image data, and to identify *evidence-path ties*, a category structurally unique to multi-image reasoning, as the most consequential source of spurious preference boundaries.

Detail ties share a conclusion but differ in granularity; binarization induces *verbosity bias* (Zheng et al., 2023), amplified in multi-image tasks where the space of describable visual details is larger. *Style/framing ties* convey the same grounded content in different discourse organization (bullets vs. prose), inducing *format preference bias*; *order-invariance ties* present the same content in a different sequence, inducing *ordering bias*.

Evidence-path ties are **unique to multi-image reasoning** and the most important for our study: two responses support the same conclusion through different image subsets $\mathcal{S}_a, \mathcal{S}_b \subseteq \mathcal{I}$ or visual cues, where (1) both reach the same conclusion, (2) $\mathcal{S}_a \neq \mathcal{S}_b$, and (3) both subsets independently provide sufficient grounding. When binarized, the model learns *evidence-path overcommitment*: it fixates on a narrow subset of images (e.g., always preferring earlier images, consistent with known position

Tie Category	Definition	Multi-image specific?	Failure mode if binarized
Detail Tie	Same conclusion, different granularity	No	Verbosity bias
Style/Framing Tie	Same content, different organization	No	Format preference bias
Evidence-Path Tie	Same conclusion via different image subsets/cues	**Yes**	Evidence-path overcommitment
Order-Invariance Tie	Same content, different presentation order	Partially	Ordering bias

Table 1: **Taxonomy of tie categories in multi-image preference data.** Each category has a distinct origin and induces a specific failure mode when forced into binary labels.

biases (Tian et al., 2025)) while ignoring equally valid evidence from others. Each failure mode is directly measurable, and Section 5.4 measures all three.

4 TieDPO: Tie-Aware Preference Optimization

4.1 Preliminaries: DPO with Ternary Preferences

Let $x = (\mathcal{I}, q)$ denote a multi-image input, where $\mathcal{I} = \{I_1, \ldots, I_n\}$ is a set of images and q a textual query. Given two candidate responses y_a and y_b, standard DPO (Rafailov et al., 2023) assumes a strict binary label. We instead consider a *ternary* preference relation $r \in \{+1,\ -1,\ 0\}$, where $r = +1$ indicates $y_a \succ y_b$, $r = -1$ indicates $y_b \succ y_a$, and $r = 0$ indicates a tie ($y_a \sim y_b$). Let π_θ denote the policy model and π_{ref} the frozen reference model, and define the *relative preference logit*:

$$\Delta_\theta(x, y_a, y_b) = \big(\log \pi_\theta(y_a \mid x) - \log \pi_\theta(y_b \mid x)\big) - \big(\log \pi_{\text{ref}}(y_a \mid x) - \log \pi_{\text{ref}}(y_b \mid x)\big). \quad (1)$$

Equation 1 is the reference-anchored quantity whose logistic likelihood defines DPO. It is well suited to strict pairs but poorly suited to measuring indifference: for an untuned model $\pi_\theta \equiv \pi_{\text{ref}}$, so $\Delta_\theta \equiv 0$ identically, and any indifference metric read off Δ_θ would then be a property of the parameterization rather than of the model. A statement about ties is a statement about what the policy itself prefers, so we define the *preference score*

$$s_\theta(x, y_a, y_b) = \beta\Big[\log \pi_\theta(y_a \mid x) - \log \pi_\theta(y_b \mid x)\Big], \qquad \log \pi_\theta(y \mid x) = \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}), \quad (2)$$

the policy's own implicit reward difference, on the same β-scale as Δ_θ so that the two are directly comparable. Throughout, Δ_θ appears only inside $\mathcal{L}_{\text{strict}}$, and every reported metric is computed from s_θ. Since s_θ is a token sum, its indifference point is length-dependent when $|y_a| \neq |y_b|$; we discuss this in Appendix D.4, control it with scale-invariant metrics (Appendix D.1), and test the unnormalized form directly (Table 17).

4.2 TieDPO Objective

For strict preference pairs ($r \neq 0$), we retain the standard DPO objective:

$$\mathcal{L}_{\text{strict}} = \log\big(1 + \exp(-\, r \cdot \beta \cdot \Delta_\theta)\big), \quad (3)$$

where $\beta > 0$ is the inverse temperature. For tie pairs ($r = 0$) we penalize separation in s_θ:

$$\mathcal{L}_{\text{tie}} = \big[\max\big(0,\ |s_\theta| - m\big)\big]^2, \quad (4)$$

with $m \geq 0$ defining an indifference zone $[-m, +m]$: inside the zone the loss vanishes, outside it a quadratic penalty pulls the score back. Since s_θ is β-scaled, m is expressed in the same units as the reward margins reported throughout.

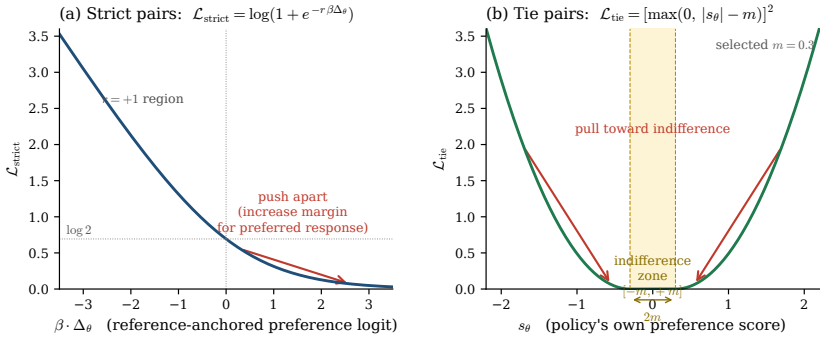


Figure 2: **Complementary loss landscape of TieDPO.** The strict and tie terms act on different quantities. (a) $\mathcal{L}_{\text{strict}}$ is a function of the reference-anchored logit Δ_θ and pushes chosen and rejected apart. (b) $\mathcal{L}_{\text{tie}}$ is a function of the policy's own preference score s_θ; it is zero within the indifference zone $[-m, +m]$ and grows quadratically outside, pulling tie-pair scores toward indifference. Selected margin $m = 0.3$.

The two terms act on different quantities. $\mathcal{L}_{\text{strict}}$ is a likelihood, and the reference term is what makes it a well-posed implicit reward model; it is unchanged from DPO. $\mathcal{L}_{\text{tie}}$ encodes a different proposition—*the policy should not prefer* y_a *over* y_b—and this proposition is about π_θ alone. Write $b(x, y_a, y_b) = \beta[\log \pi_{\text{ref}}(y_a \mid x) - \log \pi_{\text{ref}}(y_b \mid x)]$ for the reference model's own preference offset on a pair, so that $s_\theta = \Delta_\theta + b$. Penalizing $|\Delta_\theta|$ on tie pairs has fixed point $s_\theta \to b$: it drives the policy toward reproducing whichever of the two responses the reference model happened to favor, an arbitrary choice for a tied pair. Penalizing $|s_\theta|$ has fixed point $s_\theta \to 0$, which is indifference, so the quantity optimized during training matches the quantity reported at evaluation. Since $\mathcal{L}_{\text{tie}}$ constrains a difference of log-probabilities rather than their levels, a degenerate solution is possible in principle: a model could reduce $|s_\theta|$ by flattening its output distribution. The $\mathcal{L}_{\text{sft}}$ anchor opposes this, and we verify directly that it does not occur (Appendix E.4).

The total training loss combines strict preference learning, tie regularization, and a supervised fine-tuning (SFT) anchor:

$$\mathcal{L} = \underbrace{\alpha_{\text{dpo}} \cdot \mathbf{1}[r \neq 0] \cdot \mathcal{L}_{\text{strict}}}_{\text{discriminate strict pairs}} + \underbrace{\lambda_{\text{tie}} \cdot \mathbf{1}[r = 0] \cdot \mathcal{L}_{\text{tie}}}_{\text{regularize tie pairs}} + \underbrace{\lambda_{\text{sft}} \cdot \mathcal{L}_{\text{sft}}}_{\text{SFT anchor}}, \tag{5}$$

where α_{dpo} scales the strict DPO loss, λ_{tie} controls tie regularization strength, and λ_{sft} controls the SFT anchor weight. The SFT anchor is applied on the winner y_w for strict pairs and *symmetrically* averaged over y_a, y_b for tie pairs, so the anchor introduces no directional bias that would conflict with tie regularization. The indicators keep the strict and tie losses on disjoint subsets: on strict pairs TieDPO behaves identically to standard DPO; on tie pairs the margin-based penalty provides a soft indifference constraint. Together they induce a *dual dynamic*: strict pairs are pushed apart and tie pairs are pulled back into an indifference zone of width m (Figure 2; gradients in Appendix B). Maximizing the strict-pair margin is not the design target: the objective deliberately yields a smaller strict margin than DPO-strict in exchange for preserving the calibrated tie structure, a tradeoff quantified in Section 5.3.

4.3 MultiTie-22k and TieBench-MI

MultiTie-22k is built by a four-stage pipeline (cross-image question selection or generation, diverse candidate generation, informative pair construction, and ternary annotation with tie-category labels). Each tie category is paired with a source whose structure makes it naturally elicitable: Mantis storytelling for detail, NExT-QA (Xiao et al., 2021) for evidence-path, NLVR2 (Suhr et al., 2019) for order-invariance, and Mantis VQA (Jiang et al., 2024) for style/framing (Appendix A.3), with path-conditioned generation for evidence-path ties (Appendix A.4). It contains 22,198 pairs (tie rate 25.8%, Table 4) at 4.3 images per input; all four categories are represented among tie pairs (detail 21.4%, style/framing 14.6%, evidence-path 44.4%, order-invariance 19.6%).

Standard benchmarks cannot distinguish a model that recognizes indifference from one that arbitrarily picks a winner, so we build **TieBench-MI**, a 2,400-pair diagnostic benchmark in two splits. **Natural** (1,700 pairs) is drawn from unconstrained sampling over four image sources disjoint from the training data and the standard benchmarks (MMIU (Meng et al., 2024a), ReMI (Kazemi et al., 2024), Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018), Needle-in-a-Multimodal-Haystack (Wang et al., 2024c)), estimating how often ties occur and of what kind (tie rate 0.258; Table 12); **EP-Stress** (700 pairs) constructs pairs whose responses ground in disjoint sufficient image subsets, measuring behavior on evidence-path ties. Responses are generated by models disjoint from those evaluated (`Qwen3-VL235B-A22B`, `Gemini 3 Pro`); each pair is annotated in two passes, relation first and category second, by three annotators with a fourth adjudicator (Fleiss' $\kappa = 0.681$, 0.572 on tie pairs; Appendix E). Evaluation uses s_θ with threshold-free metrics defined in Section 5.3 and Appendix D.

5 Experiments

Our experiments address four questions. **(Q1)** Does tie-aware training convert multi-image preference data into downstream gains where strict-only training does not? **(Q2)** How do the objectives differ in what they learn about the tie/strict distinction, and how calibrated is that distinction? **(Q3)** Does binary preference training encode surface features—length, image position, discourse form—into the learned preference, as the taxonomy of Section 3 predicts? **(Q4)** Is each component of the objective necessary, and how sensitive is the method to its hyperparameters $(\lambda_{\text{tie}}, m)$?

5.1 Experimental Setup

Base models. We evaluate on **LLaVA-OneVision-Qwen2-7B** (Li et al., 2024a) (**LLaVA-OV**), a decoder-centric multi-image fine-tuning pipeline, and **Qwen3-VL-8B** (Bai et al., 2025), a natively multimodal model with dynamic-resolution visual encoding. Every table has two per-backbone blocks, so comparisons are always within-backbone.

Training. All models are trained with LoRA (Hu et al., 2022) (rank 128) under DeepSpeed ZeRO-3 on 4 GPUs, with $\beta = 0.1$, $\alpha_{\text{dpo}} = 1.0$ and $\lambda_{\text{sft}} = 1.0$. The tie hyperparameters $(\lambda_{\text{tie}}, m)$ are selected on a held-out validation split by the grid search of Section 5.5, which yields $\lambda_{\text{tie}} = 1.0$ and $m = 0.3$. Every reported number is the mean over three seeds $\{42, 1337, 2024\}$ (Appendix E.8); full hyperparameters are in Appendix C.

Baselines. Alongside the untuned base and **SFT**, we compare two families. *Strict-preference objectives* treat every pair as carrying a winner: **DPO-strict** trains on strict pairs only, and **MIA-DPO** (Liu et al., 2024c), the prior multi-image preference method, anchors prompts to individual images. *Tie-aware objectives* admit a third relation: **DPO-soft** assigns tie pairs target probability $\frac{1}{2}$, in the spirit of cDPO (Mitchell), and **Davidson** (Davidson, 1970) is the classical tie-extended Bradley–Terry likelihood with tie propensity ν estimated from data (Appendix E.11). Comparison within the tie-aware family is the sharper test: it isolates whether the particular form of $\mathcal{L}_{\text{tie}}$ matters, or whether any mechanism that softens tie supervision would suffice. All methods train on the same MultiTie-22k data under identical per-backbone hyperparameters. We do not compare against IPO (Azar et al., 2024) or KTO (Ethayarajh et al., 2024), which offer no mechanism for a ternary relation, nor CcDPO (Li et al., 2025), which has no public implementation.

Decontamination. MultiTie-22k is built from Mantis, NExT-QA and NLVR2, and Mantis-Eval shares a source distribution with Mantis-Instruct, so we match training images against every evaluation set and report results on the decontaminated subsets throughout (Appendix E.2).

5.2 Q1: Standard-Benchmark Performance

We evaluate on 7 standard benchmarks: 5 multi-image (**MuirBench** (Wang et al., 2024a), **MIRB** (Zhao et al., 2024), **BLINK** (Fu et al., 2024), **Mantis-Eval** (Jiang et al., 2024), **MIBench** (Liu et al., 2024b)), which probe the cross-image reasoning where evidence-path and order-invariance ties arise, and 2 single-image (**MMMU** (Yue et al., 2024), **XLRS-Bench** (Wang et al., 2025)), which check that tie-aware training does not cost single-image capability.

Model	Multi-image					Single-image		Avg
	MuirB.	MIRB	BLINK	Mantis	MIBench	MMMU	XLRS	
LLaVA-OV (Li et al., 2024a)	42.5	47.3	51.1	60.3	73.4	47.7	38.1	51.5
+ SFT	43.0	47.7	51.3	60.9	73.7	47.8	37.9	51.8
+ DPO-strict	43.7	48.3	51.9	61.4	74.1	47.9	38.0	52.2
+ MIA-DPO (Liu et al., 2024c)	43.1	48.1	52.1	61.1	74.3	47.6	38.3	52.1
+ DPO-soft	44.6	49.2	52.1	62.0	74.7	47.9	38.1	52.7
+ Davidson (Davidson, 1970)	45.1	48.7	52.5	61.6	75.0	48.0	38.2	52.7
+ **TieDPO (Ours)**	**46.6**	**50.9**	**53.9**	**63.5**	**75.8**	48.3	37.9	**53.8**
Qwen3-VL (Bai et al., 2025)	64.4	65.8	69.1	74.6	78.8	69.6	53.2	67.9
+ SFT	64.6	65.9	69.2	74.8	78.9	69.1	53.0	67.9
+ DPO-strict	64.9	66.2	69.1	74.0	79.0	69.8	53.0	68.0
+ MIA-DPO (Liu et al., 2024c)	65.1	66.3	69.5	74.2	79.1	69.5	53.3	68.1
+ DPO-soft	65.3	66.5	69.6	75.3	79.2	69.8	53.2	68.4
+ Davidson (Davidson, 1970)	65.5	65.7	69.8	74.4	79.4	69.9	53.3	68.3
+ **TieDPO (Ours)**	**66.2**	**67.3**	**70.2**	**76.0**	**79.7**	70.0	53.1	**68.9**

Table 2: **Main results on 7 standard benchmarks across both backbones.** Mantis = Mantis-Eval, XLRS = XLRS-Bench. All numbers are computed on decontaminated subsets (Appendix E.2). Avg is the unweighted mean across all 7 benchmarks. Mean over 3 seeds; standard deviations in Appendix E.8.

Findings. *(i) Tie-aware objectives outperform every strict-preference objective.* On the multi-image five-benchmark average, tie-aware methods gain 1.60–3.22 points on LLaVA-OV and 0.42–1.34 on Qwen3-VL, versus at most 0.96 / 0.30 for the strict-preference methods; DPO-strict and the tie-aware objectives see identical images and questions, so the difference isolates relation structure from data quantity. Because ties are 25.8% of MultiTie-22k, benchmark gains are modest in absolute size, and the primary evidence for our claims comes from the direct tie-pair measurements in Sections 5.3 and 5.4.

(ii) How a third relation is admitted matters. Within the tie-aware family TieDPO leads Davidson by 1.56 points on LLaVA-OV ($p = 0.047$, Table 22) and 0.92 on Qwen3-VL, with the calibration and bias measurements below providing the sharper separation. The ordering matches the prediction of Section 4: DPO-soft and Davidson anchor their tie treatment to Δ_θ (fixed point at the reference model's own arbitrary preference), whereas $\mathcal{L}_{\text{tie}}$ anchors to s_θ (fixed point at indifference).

(iii) Image anchoring does not substitute for relation structure; single-image capability is unaffected. MIA-DPO matches DPO-strict (-0.14 / $+0.20$) but trails every tie-aware objective. No method moves MMMU or XLRS-Bench by more than 0.6 points, and MultiTie-22k contains no single-image data, which licenses attributing the multi-image differences to multi-image preference structure rather than to fine-tuning in general.

5.3 Q2: Discrimination and Calibration of the Tie/Strict Distinction

Metrics. Measuring indifference invites circularity, since natural summaries of $|s_\theta|$ on tie pairs are monotone in what $\mathcal{L}_{\text{tie}}$ minimizes; every metric carrying a conclusion therefore satisfies **(R1)** non-monotonicity in the training objective and **(R2)** no free scale parameter (Appendix D.1). **Tie-vs-Strict AUC** ranks $|s_\theta|$ on strict vs. tie pairs (invariant to monotone rescaling); **Strict-Sign-Acc** is the fraction of strict pairs whose sign is correct; **Ternary Macro-F1** scores the verbalized three-way judgment over both presentation orders. **Over-confidence** is the difference between a Brier score under a calibration map frozen from the base model (*shared*) and one refit per method (*refit*), measuring how much of a method's apparent preference signal is miscalibrated rather than informative (protocol in Appendix D). We report the two TieBench-MI splits separately and never averaged (Section 4.3).

Findings. *(i) Discrimination is a precondition for the calibration claim.* The untuned models sit near chance ($\text{AUC} = 0.541$ / 0.559), and every objective improves on this, TieDPO most (0.774 / 0.781). DPO-strict reaches 0.706 / 0.714 without ever seeing a tie pair, and a length-only score reaches 0.548 (Table 8), so AUC mainly rewards a sharper reward model; it functions here as a check that a method has not destroyed the distinction it is supposed to calibrate. Strict-Sign-Acc makes

	Discrimination		Judgment	Calibration		
Model	**AUC↑**	**Sign-Acc↑**	**Macro-F1↑**	**Brier-sh.↓**	**Brier-rf.↓**	**Over-conf.↓**
Constant-`tie` / $s_\theta \equiv 0$	0.500	0.500	0.137	0.658	0.658	0.000
LLaVA-OV (Li et al., 2024a)	0.541	0.612	0.402	0.402	0.394	0.008
+ SFT	0.554	0.629	0.421	0.398	0.387	0.011
+ DPO-strict	0.706	**0.751**	0.447	0.428	0.351	0.077
+ MIA-DPO (Liu et al., 2024c)	0.692	0.734	0.438	0.421	0.359	0.062
+ DPO-soft	0.688	0.726	0.459	0.396	0.348	0.048
+ Davidson (Davidson, 1970)	0.721	0.748	0.468	0.381	0.339	0.042
+ **TieDPO (Ours)**	**0.774**	0.744	**0.509**	**0.331**	**0.312**	0.019
Qwen3-VL (Bai et al., 2025)	0.559	0.658	0.451	0.371	0.362	0.009
+ SFT	0.568	0.669	0.462	0.368	0.357	0.011
+ DPO-strict	0.714	**0.781**	0.486	0.404	0.338	0.066
+ MIA-DPO (Liu et al., 2024c)	0.698	0.767	0.479	0.397	0.343	0.054
+ DPO-soft	0.709	0.764	0.491	0.378	0.334	0.044
+ Davidson (Davidson, 1970)	0.726	0.776	0.497	0.366	0.328	0.038
+ **TieDPO (Ours)**	**0.781**	0.779	**0.534**	**0.318**	**0.301**	0.017

Table 3: **Core diagnostics on TieBench-MI-Natural.** All six metrics are threshold-free; AUC, Sign-Acc and Over-confidence are additionally invariant to monotone rescaling of $|s_\theta|$. Brier-sh. = Brier-shared (one frozen calibration map for all methods), Brier-rf. = Brier-refit (per-method map), Over-conf. = their difference. The top row gives the values a model carrying no information attains: chance discrimination, and the Brier score of predicting the class prior (class priors on this split: 0.371 / 0.258 / 0.371); a model can score worse by being confidently wrong (Table 8). EP-Stress split in Table 25; per-category breakout in Appendix E.9.

the trade explicit: DPO-strict is best (0.751 vs. our 0.744 on LLaVA-OV; 0.781 / 0.776 / 0.779 on Qwen3-VL). Tie supervision costs at most a fraction of a point of strict-pair ranking while changing the tie-side behavior substantially.

(ii) Calibration is where the objectives clearly diverge. Under the frozen shared map, DPO-strict scores worse than the untuned base (0.428 vs. 0.402 Brier on LLaVA-OV) despite its higher AUC, with over-confidence 0.077 against the base's 0.008; refitting the map recovers most of its score (0.351), which means the ranking information is present but the scale on which it is expressed is wrong. The tie-aware objectives split here: Davidson and DPO-soft reduce over-confidence to 0.042 and 0.048, better than DPO-strict but two to three times TieDPO's 0.019. This matches the fixed-point argument quantitatively: a tie penalty on Δ_θ pulls the policy toward agreeing with the reference model on tied pairs, so the resulting margins are smaller than DPO-strict's but still systematically offset, while a penalty on s_θ has no such residual. TieDPO is the only method whose over-confidence sits within roughly twice the untuned base's, so it improves discrimination without paying for it in calibration. This column, together with the surface-bias diagnostics of Section 5.4, is what our main claim rests on.

(iii) Verbalized judgment improves modestly, and per-category results are directional only. Ternary Macro-F1 rises from 0.402 to 0.509 on LLaVA-OV and 0.451 to 0.534 on Qwen3-VL, matching the calibration ordering; absolute values stay low for all methods, since preference optimization shapes a policy's relative likelihoods and only indirectly affects what the model will say when asked to adjudicate, and we make no claim of a strong multi-image judge. Table 24 breaks AUC down by tie category. The binomial intervals are wide (± 0.100 on order-invariance, $n = 140$; ± 0.162 on style ties, $n = 57$), and the direction is consistent across all four categories, with TieDPO's order-invariance advantage (0.758 vs. 0.663) the largest single gap. The EP-Stress split (Table 25) further isolates evidence-path behavior, which is too rare under unconstrained sampling (22.8% of ties, 100 pairs) for the Natural split to resolve on its own.

5.4 Q3: Surface Features in the Learned Preference

Section 3 predicts that when a tie pair is forced to yield a winner, the features available to choose by are surface features, and each tie category supplies a different one. We compute three metrics, each

paired with the category whose failure mode it tests, each on pairs annotated indifferent, and each sign-based or matched-subset and therefore invariant to monotone rescaling of s_θ (so shrinking s_θ cannot by itself improve them): **LC-Flip** (length-controlled flip rate on detail ties, with an information-free elaboration appended to the shorter response), **Position Flip Rate** (sign change under image-order reversal on order-invariance ties), and **Style Gap** (bullets-vs-prose asymmetry on style ties). Smaller is better on all three; full definitions are in Appendix E.10 and results in Table 23.

Findings. DPO-strict is worst on all three metrics and both backbones, and the mechanism is instructive: it never sees a tie pair, yet its strict-pair quality signals correlate with length and formatting, and that correlation transfers to pairs where no quality difference exists (LC-Flip 0.361 on LLaVA-OV vs. base 0.271, position flip 0.329, style gap 0.196). Discarding ties therefore trades one failure mode for another; tie supervision supplies the missing constraint, and TieDPO is the only method that falls below the base on all three metrics and both backbones. Its smaller margins do not yield unstable signs (per-$|s_\theta|$ bins in Appendix E.10), and its gap over Davidson (0.219 vs. 0.268 LC-Flip on LLaVA-OV) traces to a specific property of reference-anchoring: it pulls s_θ toward the base model's own length and position preferences rather than toward zero. Section 5.5 tests this directly by re-anchoring the tie term.

5.5 Q4: Ablations and Hyperparameter Sensitivity

Table 17 removes one term of the objective at a time and additionally replaces $\mathcal{L}_{\text{tie}}$ with a variant that penalizes $|\Delta_\theta|$ instead of $|s_\theta|$, reporting AUC and Over-confidence from TieBench-MI alongside benchmark averages.

Findings. *The reference anchor drives the result.* Replacing $|s_\theta|$ with $|\Delta_\theta|$ in $\mathcal{L}_{\text{tie}}$ costs 1.12 points of multi-image average and raises over-confidence from 0.019 to 0.048, placing this variant beside the reference-anchored baselines in Table 3 rather than beside TieDPO. With every other term held fixed, the comparison isolates which quantity is driven to zero: the reference model's arbitrary preference on a tied pair, or indifference itself. Setting $m = 0$ costs an additional 0.36 points and 0.039 AUC, so a pure quadratic pull recovers most of the benefit and the indifference zone acts as a refinement. A length-normalized variant $\bar{s}_\theta = s_\theta/|y|$ (in the manner of SimPO (Meng et al., 2024b)) is likewise slightly worse (Appendix D.4); together with TieDPO's lowest measured length sensitivity in Section 5.4, this places the mechanism on reference-anchoring rather than on length.

The surface is flat near the optimum and degrades sharply at the edges. On a 5×5 grid over $(\lambda_{\text{tie}}, m)$ (Appendix E.5), the nine cells around $(1, 0.3)$ vary by only 0.39 points and 0.028 AUC. At $\lambda_{\text{tie}} = 5$, held-out perplexity rises from 9.07 to 9.93 and mean generation length falls from 203 to 181 tokens (Table 16), the pattern expected when a policy flattens its output distribution to satisfy the tie constraint; the selected configuration sits well away from that regime. A ternary decision rule and accuracy-versus-τ sweep are in Appendix E.12.

6 Conclusion

We identified the *structured tie phenomenon* in multi-image preference data, and a DPO pathology on this data: strict-only training miscalibrates the tie/strict distinction and encodes surface features (length, image position, discourse form) into the learned preference. Across 7 benchmarks and 2 backbones, *TieDPO* (together with *MultiTie-22k* and *TieBench-MI*) is the only method that improves tie/strict discrimination without paying for it in calibration, and it also reduces all three surface biases below the untuned base. The framework extends naturally to single-image and text-only alignment, where ties also arise.

7 Limitations

Tie annotations are primarily model-generated with human verification on a subset (Fleiss' $\kappa = 0.572$ on tie labels), which bounds achievable separation on tie-side metrics (Appendix E.1). The tie hyperparameters $(\lambda_{\text{tie}}, m)$ are global; the calibration-to-downstream link is correlational (Appendix E.7); verbalized ternary judgment remains weak for all methods; and settings requiring maximum strict-pair separation may need a different operating point.

AI USE STATEMENT

In this work, we used generative AI tools in the following ways, all of which were reviewed and verified by the authors. **Data annotation and construction.** As disclosed in Section 4.3 and Appendix A, MultiTie-22k and TieBench-MI response pairs and ternary tie annotations are produced by large multimodal models (`Qwen3-VL-235B-A22B` and `Gemini 3 Pro`) with human verification on a stratified subset; annotator agreement statistics (Fleiss' κ) are reported in the main text and Appendix E. The role of generative AI here is a load-bearing methodological choice of the paper rather than a writing aid, and is analyzed accordingly (Limitations; Appendix E.1). **Writing.** We used general-purpose LLM assistants for language polishing, LaTeX formatting, and proofreading of author-written prose. All claims, derivations, results, and citations were written and verified by the authors; no scientific content was generated by AI without author verification against the underlying experiments or literature. **Code.** We used LLM coding assistants for boilerplate, LaTeX/plotting utilities, and refactoring suggestions in the training and evaluation pipeline. All code that produced reported numbers was reviewed by at least one author, and every numerical result in the tables is reproduced from logged runs. **Not used.** We did not use generative AI to conceive the research problem, propose the method, design the experiments, or generate any of the empirical results, figures, or tables. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human feedback. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Roger R Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Peter C Fishburn. Utility theory for decision making. Technical report, 1970.

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. *arXiv preprint arXiv:2501.16629*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4024–4034, 2018.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, et al. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.

Xudong Li, Mengdan Zhang, Peixian Chen, Xiawu Zheng, Yan Zhang, Jingyuan Zheng, Yunhang Shen, Ke Li, Chaoyou Fu, Xing Sun, et al. Zooming from context to cue: Hierarchical preference optimization for multi-image mllms. *arXiv preprint arXiv:2505.22396*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22417–22428, 2024b.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*, 2024c.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference alignment of llms via optimal transport. *Advances in Neural Information Processing Systems*, 37:104412–104442, 2024.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024a.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024b.

Eric Mitchell. A note on dpo with noisy preferences and relationship to ipo, 2023. *URL https://ericmitchell. ai/cdpo. pdf*.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Pejaver V Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6418–6428, 2019.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10599–10609, 2025.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14325–14336, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024c.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2(3):8, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 37:26171–26194, 2024.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

A Data Construction Details

A.1 MultiTie-22k Dataset Statistics

Statistic	Value
Total response pairs	22,198
Strict pairs ($r \neq 0$)	74.2%
Tie pairs ($r = 0$)	25.8%
Tie breakdown (share of tie pairs):	
Detail ties	21.4%
Style/framing ties	14.6%
Evidence-path ties	44.4%
Order-invariance ties	19.6%
Avg. images per input	4.3
Avg. response length (tokens)	137

Table 4: **MultiTie-22k dataset statistics.** Summary numbers referenced in Section 4.3; per-category data sources and construction details follow in the rest of this appendix.

A.2 Overall Pipeline

Our data construction follows a four-stage pipeline:

1. **Question selection/generation.** For each multi-image input $\mathcal{I}$, we select or generate a question q that requires reasoning across multiple images (e.g., comparison, sequential reasoning, or aggregation).
2. **Candidate response generation.** We generate K diverse candidate responses using a strong multimodal model with varied prompting strategies (e.g., different system prompts encouraging different evidence paths, detail levels, or organizational styles).
3. **Response pair construction.** We form informative pairs from the K candidates, prioritizing pairs that are likely to be close in quality (to maximize the yield of tie pairs) as well as pairs with clear quality differences (to ensure sufficient strict pairs).
4. **Preference annotation.** Each pair is annotated with a ternary label (`better_a`, `better_b`, or `tie`) and, for tie pairs, a tie category from our taxonomy.

All question generation, candidate response generation, and pairwise preference judging calls are dispatched to **Claude 3.7 Sonnet** through an OpenAI-compatible chat-completions endpoint. We

use a shared infrastructure layer across all four per-category pipelines that handles: (i) thread-local OpenAI clients with a global token-bucket rate limiter (2 calls/second) and exponential-backoff retries for rate-limit and transient errors; (ii) image preprocessing that resizes each image so that its total pixel count does not exceed 768×768, converts it to RGB JPEG (quality 85), and encodes it as a base64 `data:image` URL; (iii) a JSON-only system prompt together with a fault-tolerant parser that recovers from common wrapping patterns (code fences, trailing text) by scanning for the outermost balanced JSON brackets; and (iv) multi-threaded sample-level processing (`MAX_WORKERS=4`) with per-sample JSONL writes guarded by a writer lock so that partial progress is always persisted. Temperature is chosen per call type: 0.2 for question generation (to keep the core-fact structure stable), 0.7–0.9 for candidate response generation (to encourage diversity across detail levels, evidence paths, orderings, and styles), and 0.1 for pairwise judging (to stabilize labels). Complete prompt templates and runner scripts are included in the supplementary material.

A.3 Source-to-Category Mapping

Rather than drawing all four tie categories from a single homogeneous corpus, we pair each category with a multi-image dataset whose structural properties make that category naturally elicitable (Table 5). This design keeps each per-category pipeline focused: the prompts used for one category can explicitly exclude the other three (e.g., the detail-tie prompts explicitly forbid organization, sufficiency, and evidence-path variation), making the resulting tie labels cleaner and the tie-category annotation more reliable.

Source	Tie Category	Why this source fits
Mantis visual storytelling (Jiang et al., 2024)	Detail	Multi-image narrative descriptions naturally admit answers at different granularities while still covering the same core events.
NExT-QA (Xiao et al., 2021)	Evidence-path	Multi-frame video QA allows the same conclusion to be supported by distinct frame subsets (different temporal segments or visual cues).
NLVR2 (Suhr et al., 2019)	Order-invariance	Two-image reasoning tasks whose correctness is invariant to the order in which the two images are discussed.
Mantis multi-image VQA (Jiang et al., 2024)	Style/framing	Open-ended multi-image QA permits multiple organizationally distinct (bullet vs. prose, conclusion-first vs. evidence-first) yet semantically equivalent answers.

Table 5: **Per-category source mapping for MultiTie-22k.**

A.4 Per-Category Generation Pipelines

All four per-category pipelines share the same four-stage skeleton (Appendix A.4.1); they differ only in (i) the underlying multi-image source, (ii) the *variation axis* along which positive candidates are generated, and (iii) the judge-level checks that enforce category purity. We first describe the shared skeleton once, then for each category describe only what is specialized, and provide one worked example per category (Figures 3–6) illustrating a representative tie pair.

A.4.1 Shared Four-Stage Skeleton

Let a sample consist of an image list $\mathcal{I}$ (optionally with ordering or frame indices) and, if available, an original dataset conversation or label. Every per-category pipeline proceeds through the same four stages; category-specific deviations are tabulated in Table 6.

Stage 1: Question proposal. We prompt Claude 3.7 Sonnet in *multimodal* mode to produce a question that both requires reasoning across multiple images in $\mathcal{I}$ and is structurally suited to eliciting the target tie category. The prompt explicitly defines the target tie category, lists category-friendly question templates, and excludes question types that would foreclose the target variation axis (e.g., detail-tie questions exclude fixed-length answers; evidence-path-tie questions exclude questions answerable from a single image). The model returns a structured JSON object containing the question together with two control fields used throughout the rest of the pipeline: `core_facts`—a 2–5 item

list of facts every correct answer *must* include—and `optional_elaborations`—a 3–6 item list of supporting details that *may* be added. This core/optional decomposition is what later lets candidates vary along exactly one axis without turning a tie into a different category (in particular, without turning it into a sufficiency tie).

Stage 2: Category-specific candidate generation. Conditioned on the question and its `core_facts`/`optional_elaborations`, we generate K correct candidate responses that vary along the *single axis* associated with the target tie category (granularity for detail, organization for style, evidence subset for evidence-path, presentation order for order-invariance; see Table 6). All other axes are held fixed by hard in-prompt constraints: every response must cover *all* `core_facts`; optional elaborations must be drawn only from `optional_elaborations` (not invented); and axes not being varied are explicitly pinned (e.g., "all responses must use the same organization"). These constraints prevent leakage across tie categories at generation time and are what make the later judge-level category-purity checks tractable.

Stage 3: Flawed-variant generation (for strict pairs). In parallel we generate 4 flawed candidates drawn from the error taxonomy {`factual_error`, `hallucinated_detail`, `wrong_conclusion`, `missing_critical_point`, `misattributed_evidence`}. These serve as the rejected side in *strict* DPO pairs. Flawed responses are generated at varying detail levels so that each can later be matched with a correct response of similar length, preventing strict DPO pairs from being decided by length alone—a common form of verbosity bias in DPO training data.

Stage 4: Pair construction and ternary judging. Tie-candidate pairs are formed by selecting two correct responses whose variation along the target axis is maximal (most diagnostic) while all other axes match exactly. Strict DPO pairs match each flawed response with the correct response whose detail level is closest to it. Every candidate pair is sent to Claude 3.7 Sonnet and receives a ternary label together with a structured rationale. Each category's judge prompt encodes a short sequential checklist (e.g., same conclusion, both cover all core facts, both grounded, differ mainly along the target axis, all other axes match); the judge refuses a tie label if any item fails and refuses a strict label when the difference is attributable to length or style alone. Pairs with `is_strict_preference_reliable=no` or that fail any purity check are discarded rather than reassigned to a neighboring category, so each split stays clean by construction.

Category	Source	Stage-2 variation axis	Stage-4 purity checks
Detail	Mantis visual storytelling (Jiang et al., 2024)	Word count across 5 detail levels, from ∼25–45 to ∼150–200 words	Same conclusion, same core facts, same evidence, same organization, differ mainly in detail level
Evidence-Path	NExT-QA (Xiao et al., 2021)	Image (frame) subset $\mathcal{S}_a$ vs. $\mathcal{S}_b$, with $\lvert\mathcal{S}_a \cap \mathcal{S}_b\rvert \leq 1$	Same conclusion, both subsets independently sufficient, neither response uses evidence outside its subset
Order-Invariance	NLVR2 (Suhr et al., 2019)	Discussion order of I_1 and I_2 (two responses: $I_1 \to I_2$ vs. $I_2 \to I_1$)	Same conclusion, same content coverage, same organization and detail level, differ only in presentation sequence
Style/Framing	Mantis multi-image VQA (Jiang et al., 2024)	Organizational form: evidence-first paragraph vs. conclusion-first bullets	Same conclusion, same core-fact coverage, same evidence path, similar detail level, differ only in organization

Table 6: **Per-category specializations of the shared four-stage skeleton.** Each row defines the only pipeline elements that change; all other machinery (question proposal, flawed-variant generation, judge-prompt structure, rate limiting, JSON parsing, persistence) is shared.

A.4.2 Detail Tie Pipeline (Visual Storytelling)

Multi-image narrative descriptions naturally admit answers at different granularities while still covering the same core events, which makes Mantis visual storytelling (Jiang et al., 2024) the natural source for detail ties. We follow the shared skeleton with the specializations in Table 6. In Stage 1 we

sample up to six images per story while preserving temporal order, and bias question generation toward templates such as "Tell me the story these images show" that structurally admit responses of varying length. In Stage 2 we generate five candidates $r_1, \ldots, r_5$ whose target word counts scale roughly geometrically from ∼25–45 (concise) to ∼150–200 (elaborate+), with organization (paragraph) and register (narrative-descriptive) pinned to identical values across all five. In Stage 4, tie pairs are formed by selecting two candidates whose detail-level gap is ≥ 1 and whose `organization` and `register` fields match exactly; the largest detail gaps are prioritized. The Stage-2 prompt enforces category purity in-prompt via an explicit "what to avoid" list that forbids switching organization form, changing discourse order, skipping any core fact, or relying on different images; this excerpt is reproduced below.

```
Target tie type: Detail Tie ONLY.
Same conclusion, same core facts, same evidence, SIMILAR organization,
DIFFERENT level of elaboration.

What to avoid (these would make it a DIFFERENT tie type):
- Switching organization form (bullet vs paragraph)
  -> style tie.
- Changing discourse order (conclusion-first vs evidence-first)
  -> style tie.
- Skipping any core fact in some responses but not others
  -> sufficiency tie.
- Relying on different images / different visual cues
  -> evidence-path tie.
```

Example: Detail Tie *Source: Mantis visual storytelling*

Images ($\mathcal{I}$, 6 total, 3 shown):





...

Question (q): What notable activities and recognitions took place at this indoor July 4th barbecue gathering?

Core facts: indoor July 4th barbecue; guests socialize and share a meal; a chef presents food; recognition awards are given to attendees.

Response A (concise, ∼30 words):
At this indoor July 4th barbecue gathering, guests gathered to socialize and enjoy a meal, a chef was present to present food, and recognition awards were given to attendees.

Response B (elaborate, ∼75 words):
At this indoor July 4th barbecue gathering, guests gathered to socialize—with some playfully interacting while others ate and conversed in large groups—and enjoy a meal; a chef presented food with both hands raised in a playful gesture, and recognition awards were given to attendees by a pair of people wearing red lanyards.

Label: `tie` (Detail). **Why this is a detail tie:** both responses cover the identical core facts using the same paragraph organization; Response B only adds optional descriptive elaboration (playful interactions, the chef's gesture, the red lanyards) around the same grounded events.

Figure 3: **Worked example of a Detail Tie pair.** The two responses cover the identical set of core facts and use the same paragraph organization; they differ only in how much optional elaboration is included.

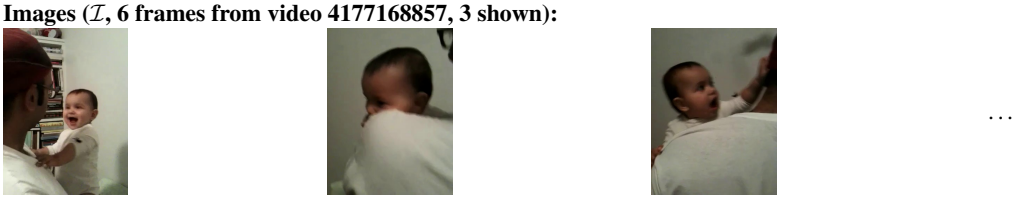
A.4.3 Evidence-Path Tie Pipeline (NExT-QA)

NExT-QA's multi-frame video QA structure is what makes evidence-path ties harvestable at scale: for many questions the same conclusion can be supported independently by distinct temporal segments or visual-cue subsets of the video. We follow the shared skeleton with two category-specific modifications beyond the specializations in Table 6.

(i) Evidence-path identification. In Stage 1, after question proposal, we additionally identify at least two frame subsets $\mathcal{S}_a, \mathcal{S}_b \subseteq \mathcal{I}$ with $\mathcal{S}_a \neq \mathcal{S}_b$ and $|\mathcal{S}_a \cap \mathcal{S}_b| \leq 1$ that each independently suffice to answer the question; questions that cannot be decomposed into at least two disjoint sufficient subsets are filtered out at this step. *(ii) Path-conditioned generation.* In Stage 2, each candidate response is generated under a *subset-masked* context that exposes only the frames in its assigned subset; this prevents accidental mixing of evidence paths during generation rather than having to catch it at judging time. Stage 4 additionally runs a full-image grounding validation, verifying that neither response makes claims outside its assigned subset and that both subsets are still independently sufficient in the presence of the full image set. We also retain evidence-path ties that arise *organically* during unconstrained candidate generation in the other three pipelines; these pass through the same Stage 4 verification. In our final dataset, 17% of evidence-path tie pairs are organically occurring and 83% are constructed via path-conditioned generation; results in Section 5 hold on both subsets.

Example: Evidence-Path Tie *Source: NExT-QA*

Images ($\mathcal{I}$, 6 frames from video 4177168857, 3 shown):



...

Question (q): What consistent behaviors show that the baby is enjoying playing with the man?

Evidence subsets: $\mathcal{S}_a = \{I_0, I_4\}$ (facial expressions: open-mouthed laugh, clinging contact), $\mathcal{S}_b = \{I_1, I_6\}$ (physical interactions: reaching for the man's head, gripping his shirt); both independently sufficient.

Response A (grounds in $\mathcal{S}_a$):
The baby shows enjoyment through having an open-mouthed, joyful laugh with an excited face in one moment, then clinging closely to the man in comfortable, engaged physical contact—these consistent behaviors confirm the baby is enjoying playing with him.

Response B (grounds in $\mathcal{S}_b$):
The baby's playful physical interactions show they're enjoying playing with the man: the baby playfully reaches up to grab the man's head/hair, and later holds tightly onto the man's shirt during close interaction, both actions demonstrating engaged, joyful play.

Label: `tie` (Evidence-Path). **Why this is an EP tie:** both responses reach the same conclusion (the baby is enjoying the play) but cite disjoint frame subsets—A cites facial/emotional cues, B cites motor/contact cues—and each subset is independently sufficient evidence.

Figure 4: **Worked example of an Evidence-Path Tie pair.** Responses A and B reach the same conclusion using disjoint frame subsets; forcing a binary preference would implicitly teach the model to prefer one evidence path over an equally valid alternative.

A.4.4 Order-Invariance Tie Pipeline (NLVR2)

NLVR2 consists of two-image reasoning pairs (I_1, I_2) paired with a textual statement whose truth value depends on both images; the task guarantees that, conditional on the conclusion being correct, the order in which I_1 and I_2 are discussed does not affect correctness. We follow the shared skeleton with two small specializations. In Stage 1 we reformulate the NLVR2 verification statement into an open-ended "explain whether and why" question that requires referring to both images. In Stage 2 we generate *exactly two* correct candidates under matched prompts that differ only in a single instruction pinning discussion order ($I_1 \to I_2$ vs. $I_2 \to I_1$); organization (paragraph), register, target detail level, and required evidence points are held identical across the two prompts. In Stage 4 the judge is asked to confirm four conditions simultaneously—same conclusion, same content coverage, same organization and detail level, difference *only* in presentation sequence—and any pair failing any condition is discarded rather than reassigned, keeping the split clean.

Example: Order-Invariance Tie *Source: NLVR2*

Images ($\mathcal{I}$, 2 total):

I_1: 

I_2: 

Question (q): Is it true that an image shows one leather pencil case, displayed open with writing implements tucked inside?

Response A (discusses I_1 before I_2):
Looking at the left image, I see a brown leather case that is open and has elastic loops inside to hold writing implements, but it's clearly a pen case or writing instrument holder, not a pencil case specifically. The right image shows several rolled leather cases in different colors (red, purple, brown, black), all closed. None of these images shows a single leather pencil case displayed open with writing implements inside. Therefore, the statement is False.

Response B (discusses I_2 before I_1):
Examining the right image first, I see several rolled leather cases in different colors (red, purple, brown, black) that are all closed, not open. The left image shows a brown leather case that is unzipped and open, revealing elastic loops designed to hold pens or other writing instruments. While this is a leather case for writing implements, neither image shows what could specifically be called a pencil case with writing implements tucked inside. Therefore, the statement is False.

Label: `tie` (Order-Invariance). **Why this is an order tie:** both responses reach the identical *False* verdict with the same supporting observations and the same paragraph organization; they differ only in whether I_1 or I_2 is described first.

Figure 5: **Worked example of an Order-Invariance Tie pair.** The two responses contain the same information and reach the same conclusion but sequence the two images in opposite orders.

A.4.5 Style/Framing Tie Pipeline (Mantis Multi-Image VQA)

Mantis's multi-image VQA subset (Jiang et al., 2024) contains open-ended cross-image questions whose acceptable answer space spans multiple organizational forms. We follow the shared skeleton with two specializations. In Stage 1 we additionally require a fixed target detail level (medium, $\sim$70–100 words) to remove granularity as a source of variation. In Stage 2 we generate exactly two correct candidates under identical content constraints but opposite stylistic instructions: (a) *evidence-first paragraph*—prose form that states evidence across images before concluding; and (b) *conclusion-first bullets*—states the answer up front, then enumerates supporting evidence as a bullet list. In Stage 4 the judge verifies same conclusion, same core-facts coverage, same evidence path, and similar detail level; only the organizational form should differ, and any pair with differences beyond organization is discarded.

A.5 Annotation Protocol

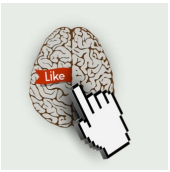
Preference annotation combines API-based model annotation and human verification:

- **Primary annotation.** All candidate generation and pairwise judging is performed with Claude 3.7 Sonnet. Question generation uses temperature 0.2; candidate generation uses 0.7–0.9 (by pipeline); judging uses 0.1.
- **Structured output.** Every judge call returns not just a ternary label but a structured rationale including category-specific check fields (e.g., `same_conclusion`, `both_cover_all_core_facts`, `differ_mainly_in_detail_level` for detail-tie judging). This enables fine-grained filtering beyond the label itself.
- **Human verification.** A random subset of 30% of annotations is verified by human annotators against the same per-category criteria used by the judge.
- **Quality filtering.** Pairs with low-confidence annotations, fallback-parsed outputs, or `is_strict_preference_reliable=no` are re-annotated or discarded.

Example: Style/Framing Tie *Source: Mantis multi-image VQA*

Images ($\mathcal{I}$, 2 total):





Question (q): How could Walgreens Pharmacy integrate digital engagement concepts like the "Like" button into their physical store experience?

Core facts: Walgreens storefront and a large "Like" button serve as visual cues; integration can use interactive kiosks / in-store tablets / app ratings; benefits include customer feedback and stronger engagement between physical and digital channels.

Response A (evidence-first paragraph):
Walgreens Pharmacy could enhance customer experience by incorporating digital "Like" button concepts into their physical stores. Interactive kiosks could allow customers to rate their shopping experience or specific products, similar to clicking a digital "Like" button. This integration would benefit the store by gathering valuable customer feedback while making shoppers feel more engaged with the brand, bridging the gap between traditional retail and digital interaction.

Response B (conclusion-first bullets):
Creating a seamless blend of digital engagement within Walgreens' physical space would enhance the customer experience significantly. Consider:

- In-store tablets for instant product feedback (like the digital "Like" button shown)
- App integration allowing customers to rate their pharmacy visit before leaving

This approach would generate valuable customer insights while making shoppers feel their opinions matter—just as they do in the digital world.

Label: `tie` (Style/Framing). **Why this is a style tie:** identical conclusion and identical core-fact coverage at a similar detail level; the only substantive difference is discourse form (prose with evidence-first vs. bullets with conclusion-first).

Figure 6: **Worked example of a Style/Framing Tie pair.** The two responses convey identical grounded content but differ in discourse structure (prose with evidence-first vs. bullets with conclusion-first).

A.6 Robustness Considerations

Three implementation details improve pipeline robustness at scale. *(i) Rate-limit-aware retries.* API calls are retried up to six times with a shared global rate limiter; rate-limit errors trigger exponential backoff (base 10s, doubling, capped at 120s) with jitter, while other transient errors use linear backoff. *(ii) Fault-tolerant JSON parsing.* Judge and generation outputs occasionally wrap valid JSON in code fences or add trailing commentary; our parser first attempts a direct `json.loads`, then strips code-fence markers, then scans for the outermost balanced JSON brackets. Unparseable outputs fall back to a structured default record marked with `confidence=low` and `is_strict_preference_reliable=no` so that downstream filtering can remove them. *(iii) Per-sample persistence and error records.* Every sample's question record, tie-pair records, and strict-pair records are written to three separate JSONL files immediately upon completion, guarded by a writer lock. Failed samples (missing images, repeated API failures) produce a structured error record rather than halting the pipeline, preserving partial progress.

B Gradient Analysis

To build intuition for how TieDPO shapes the preference landscape, we analyze the gradients of the strict and tie losses.

Strict pairs: push apart. For a strict pair with $r = +1$:

$$\frac{\partial \mathcal{L}_{\text{strict}}}{\partial \Delta_\theta} = -\beta \cdot \sigma(-\beta \cdot \Delta_\theta), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function. This gradient is always negative, pushing Δ_θ toward $+\infty$. The magnitude is largest when $\Delta_\theta \approx 0$ and vanishes as $\Delta_\theta \to +\infty$.

Tie pairs: pull together. For a tie pair:

$$\frac{\partial \mathcal{L}_{\text{tie}}}{\partial s_\theta} = \begin{cases} 2 \cdot \text{sign}(s_\theta) \cdot (|s_\theta| - m), & \text{if } |s_\theta| > m, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

When the preference score exceeds the margin, the gradient pushes s_θ back toward the indifference zone; within the zone, the gradient is zero. Because s_θ involves π_θ alone, this gradient flows through the policy with no reference-model term—which is why the fixed point of $\mathcal{L}_{\text{tie}}$ is indifference in the policy rather than agreement with the reference model (Section 4).

Dual dynamics. Together, these two forces create a *dual dynamic*: strict pairs are pushed apart to maintain discrimination, while tie pairs are pulled together to preserve indifference (Figure 2 in the main text).

C Implementation Details

Hyperparameter	Value
Base models	LLaVA-OneVision-Qwen2-7B, Qwen3-VL-8B
LoRA rank	128
LoRA alpha	256
Learning rate	2e-5
Batch size (per GPU)	1
Gradient accumulation steps	8
Number of GPUs	4
DeepSpeed stage	ZeRO-3
Training epochs	3
Warmup ratio	0.03
β (DPO temperature)	0.1
α_{dpo} (strict DPO loss weight)	1.0
λ_{tie} (tie regularization weight)	1.0
m (tie margin)	0.3
λ_{sft} (SFT anchor weight)	1.0

Table 7: **Training hyperparameters.**

Compute budget. All TieDPO training runs use 4×H20 GPUs. A single TieDPO training run on MultiTie-22k takes approximately 1.5 GPU-hours for LLaVA-OV and 4 GPU-hours for Qwen3-VL; the full study—three seeds per configuration, the 5×5 hyperparameter grid, and the ablations—totals on the order of a few hundred GPU-hours, and inference across the 7 standard benchmarks plus TieBench-MI adds roughly 10 GPU-hours.

Evaluation infrastructure. All benchmark evaluations are conducted using the `lmms-eval` framework. LoRA checkpoints are merged into full HuggingFace models before evaluation. Inference is run with batch size 1 on 4 GPUs using `accelerate`.

D TieBench-MI Metrics: Definitions and Protocols

D.1 Design Requirements

Evaluating indifference differs from evaluating accuracy in one respect that shapes everything below: the most obvious summary statistics are also the ones a tie-regularized model optimizes directly. If a metric counts how often $|s_\theta|$ falls below a threshold, or averages $|s_\theta|$ over tie pairs, then a model trained to shrink $|s_\theta|$ improves on it without having learned to distinguish tied pairs from strict ones. In the limit, $s_\theta \equiv 0$ is optimal.

Every metric that carries a conclusion in this paper therefore satisfies two requirements.

(R1) Not a monotone transform of the training objective. A metric must be capable of decreasing when $\mathcal{L}_{\text{tie}}$ decreases. Tie-vs-Strict AUC satisfies this because it depends on the *separation* between the tie and strict score distributions: uniformly shrinking $|s_\theta|$ shrinks both and leaves AUC unchanged, while shrinking only the tie population raises it and shrinking only the strict population lowers it. Over-confidence satisfies it because both of its terms decrease under useful sharpening and only the first under useless sharpening.

(R2) No free scale parameter. A metric must not depend on a constant chosen by us. Our earlier design iterations included threshold-indexed quantities, and the ordering of methods under them moved with the threshold; such a metric reports the analyst's choice as much as the model's behavior. AUC ranges over all thresholds, Sign-Acc needs none, Macro-F1 reads a label the model emits directly, and the Brier pair fixes its calibration map by a stated protocol rather than by selection.

Quantities that fail either requirement are still informative as descriptions of model behavior, and we report several in Appendix D.5. They are not used to order methods.

D.2 Metric Definitions

Let $\mathcal{D} = \{(x_i, y_{a,i}, y_{b,i}, r_i)\}$ with $r_i \in \{-1, 0, +1\}$, and write $s_i = s_\theta(x_i, y_{a,i}, y_{b,i})$ from Eq. 2. Let $\mathcal{T} = \{i : r_i = 0\}$ and $\mathcal{S} = \{i : r_i \neq 0\}$.

Tie-vs-Strict AUC.

$$\text{AUC} = \frac{1}{|\mathcal{S}||\mathcal{T}|} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{T}} \Big[\mathbb{1}\big(|s_i| > |s_j|\big) + \tfrac{1}{2}\, \mathbb{1}\big(|s_i| = |s_j|\big) \Big]. \tag{8}$$

This is the rank statistic of $|s_\theta|$ against the binary tie/strict label. It is invariant under any strictly increasing map applied to $|s_\theta|$, so multiplying all margins by a constant—the effect of temperature-like sharpening—leaves it unchanged. A model that has learned nothing about the distinction scores 0.5, and a model whose tie pairs are *more* separated than its strict pairs scores below 0.5, which makes the metric two-sided and therefore falsifying.

Strict-Sign-Acc.

$$\text{Sign-Acc} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbb{1}\big(\text{sign}(s_i) = r_i\big), \tag{9}$$

included so that a method cannot improve its tie behavior by degrading its ability to rank genuinely unequal responses. Chance is 0.5.

Ternary Macro-F1. Each pair is presented to the model with an instruction to answer $\{\texttt{A better}, \texttt{B better}, \texttt{equivalent}\}$, in both presentation orders. A pair counts as correct only if both orders yield the label consistent with the ground-truth relation under that order, so position sensitivity registers as error rather than cancelling between the two presentations. We report the unweighted mean of per-class F1, which prevents the majority class from dominating.

Ternary Brier scores. A calibration map $g_\phi : \mathbb{R} \to \Delta^3$ turns a scalar score into a distribution over $\{-1, 0, +1\}$. We use an ordinal logistic form with two cutpoints,

$$g_\phi(s) = \big(\sigma(c_1 - s/T),\ \sigma(c_2 - s/T) - \sigma(c_1 - s/T),\ 1 - \sigma(c_2 - s/T)\big), \tag{10}$$

$\phi = (c_1, c_2, T)$ fit by maximum likelihood, and score

$$\text{Brier} = \frac{1}{|\mathcal{D}|} \sum_i \sum_{k \in \{-1,0,+1\}} \big(g_\phi(s_i)_k - \mathbb{1}(r_i = k)\big)^2. \tag{11}$$

D.3 Why Two Calibration Protocols

The two Brier variants differ only in where ϕ comes from, and the pair is more informative than either alone.

Score function	AUC	Sign-Acc	Brier-sh.	Brier-rf.	Over-conf.
$s_\theta \equiv 0$	0.500	0.500	0.658	0.658	0.000
Random $s_\theta \sim \mathcal{N}(0,1)$	0.500	0.500	0.703	0.661	0.042
Length-only $s_\theta = \lvert y_a\rvert - \lvert y_b\rvert$	0.548	0.521	0.681	0.643	0.038
Oracle relation	1.000	1.000	0.000	0.000	0.000

Table 8: **Reference values for degenerate and surface-feature score functions** on TieBench-MI-Natural. Rows with analytic values are given exactly; the length-only predictor is measured, and indicates how much of the tie/strict distinction is recoverable from response length alone, i.e. the floor above which a learned score must rise to be interesting. All analytic values use the empirical class distribution of the split (0.371 / 0.258 / 0.371), not a uniform prior. Note that the random score is *worse* than $s_\theta \equiv 0$ on Brier-shared (0.703 vs. 0.658): degenerate predictors do not upper-bound the Brier columns, since confident wrong predictions cost more than uninformative ones.

Brier-refit fits ϕ separately for each method on that method's own validation scores. Since the map absorbs any monotone rescaling, this measures discriminative content and nothing else: two methods whose scores are related by a monotone transform receive identical Brier-refit. It answers whether a method's scores contain the information needed to predict the relation.

Brier-shared fits ϕ once, on the base model's validation split, and applies that frozen map to every method. A method whose score distribution has drifted from the base model's—systematically wider margins, say—now receives probabilities that do not match its empirical accuracies, and pays for it.

Fitting per method only, which is the more common practice, would defeat the purpose here. Calibration of the tie/strict distinction is part of what we claim to improve, and a per-method map removes exactly that quantity before scoring. We therefore report both and take their difference, Over-confidence, as the direct measure: it is zero when a method's scores are calibrated on the base model's scale, and grows with the mismatch. It is bounded below by zero because refitting can only improve fit on the same data. Reference values for degenerate score functions appear in Table 8, so that the achievable range of each metric is visible.

D.4 Length Dependence of the Preference Score

The preference score s_θ is a token sum, and detail ties differ in length by construction, so the length dependence of Eq. 2 deserves an explicit treatment. When $|y_a| \neq |y_b|$, the condition $s_\theta = 0$ does not say the two responses are equally likely per token: it says their total log-probabilities coincide, which requires the longer response to have the *higher* average per-token likelihood. The indifference point that $\mathcal{L}_{\text{tie}}$ targets is therefore length-dependent.

This property is inherited from the DPO reward parameterization, not introduced by the tie term. The implicit reward $\beta \log \pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)$ is defined on the same unnormalized token sum, and $\mathcal{L}_{\text{strict}}$ already ranks responses by it. Normalizing only inside $\mathcal{L}_{\text{tie}}$ would therefore make the two terms act on different scales: the strict term would push apart an unnormalized difference while the tie term pulled together a per-token average, and the fixed points of the two would no longer be commensurable. The consistency requirement is what selects the unnormalized form—the tie term should constrain the same quantity the strict term ranks and the same quantity the reward is defined on—not a preference about which scale is intrinsically better behaved.

The requirement is falsifiable, and we test it rather than assert it: applying $\mathcal{L}_{\text{tie}}$ to $\bar{s}_\theta = s_\theta/|y|$ is one row of Table 17, and it is slightly worse than the unnormalized form on both backbones rather than better (Section 5.5). Length dependence is thus not what drives the result. Beyond that we control it where it could distort a conclusion: every pair is scored in both presentation orders, the reported metrics are invariant to monotone rescaling of $|s_\theta|$ (Appendix D.1), and the length-controlled flip rate of Section 5.4 measures the residual length sensitivity of each method on a length-matched subset with a content-preserving intervention, rather than assuming it away. We deliberately do not measure this with a correlation between s_θ and length, since $\mathcal{L}_{\text{tie}}$ shrinks s_θ on exactly the pairs such a correlation would be computed over, making it partly a restatement of the training objective.

Model	Mean $\lvert s_\theta \rvert$			Tie-pair quantiles			
	tie	**strict**	**ratio**	p_{50}	p_{90}	p_{99}	$\overline{\lvert y \rvert}$
LLaVA-OV (Li et al., 2024a)	0.37	0.42	1.14	0.29	0.78	1.42	214
+ DPO-strict	0.58	0.94	1.62	0.44	1.14	2.03	231
+ DPO-soft	0.36	0.66	1.83	0.27	0.71	1.34	212
+ Davidson (Davidson, 1970)	0.33	0.64	1.94	0.25	0.67	1.28	209
+ TieDPO (Ours)	0.19	0.57	3.00	0.15	0.41	0.86	203

Table 9: **Score-distribution statistics on TieBench-MI-Natural**, all in β-scaled reward units. Descriptive only: the mean absolute score on tie pairs is what $\mathcal{L}_{\text{tie}}$ minimizes, so a low value for TieDPO is expected and carries no evidential weight. The ratio column is informative in a way the absolute columns are not, and is the quantity AUC formalizes without depending on distributional shape. The tie-pair quantiles of the untuned base model are what fix the search range for m in Section 5.5; mean response length $\overline{\lvert y \rvert}$ in tokens is included because s_θ is a token sum, so the absolute columns are not comparable across rows with different generation lengths.

Model	0.02	0.05	0.1	0.2	0.3	0.5	0.8	1.2	2.0	3.0
LLaVA-OV (Li et al., 2024a)	0.455	0.456	0.461	0.471	0.481	0.495	0.490	0.448	0.345	0.280
+ DPO-strict	0.557	0.558	0.560	0.563	0.565	0.564	0.542	0.486	0.363	**0.285**
+ DPO-soft	0.541	0.546	0.559	0.588	0.613	0.633	0.599	0.508	0.363	0.285
+ Davidson (Davidson, 1970)	0.557	0.564	0.580	0.614	0.642	0.660	0.615	**0.517**	**0.366**	0.285
+ TieDPO (Ours)	**0.560**	**0.581**	**0.624**	**0.695**	**0.724**	**0.704**	**0.623**	0.516	0.365	0.285
All-`tie` rule	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258

Table 10: **Ternary accuracy at each decision threshold** τ, TieBench-MI-Natural, LLaVA-OV, mean over 3 seeds. The majority-class row is the accuracy of always predicting the most frequent relation, and marks the level at which a method's accuracy stops being informative.

D.5 Descriptive Statistics

The following quantities describe how the score distributions move under each objective. They are threshold-dependent or objective-aligned, so they fail R1 or R2 and we do not use them to rank methods; they are reported because they make the mechanism behind the core metrics concrete.

D.6 Threshold Curve: Numerical Values

Table 10 gives the values plotted in Figure 7. The grid spans from $\tau = 0.02$, below which almost no pair is called a tie, to $\tau = 3$, above which almost every pair is, so the two degenerate ends of the decision rule bracket the reported range.

E TieBench-MI Construction

E.1 Sources, Generation, and Annotation

TieBench-MI is built so that four aspects of its construction vary independently: the image source, the response generator, the annotation procedure, and whether the annotator knows which tie category a pair was drawn for. Holding these separate is what allows a result on the benchmark to be attributed to model behavior rather than to a property of how the data was made.

Image sources. All images come from **MMIU** (Meng et al., 2024a), **ReMI** (Kazemi et al., 2024), **Spot-the-Diff** (Jhamtani & Berg-Kirkpatrick, 2018), and **Needle-in-a-Multimodal-Haystack** (Wang et al., 2024c). These are disjoint from the three sources of MultiTie-22k and from all 7 benchmarks in Table 2, so no evaluation image has been seen in training by any method, and no benchmark result is reused as a diagnostic. The four sources differ in the cross-image relation they emphasize—semantic grouping, referential resolution, fine-grained difference, and retrieval under distraction—which is what makes per-source breakouts meaningful.

Image source	#Pairs	#Natural	#EP-Stress	Tie rate (Nat.)
MMIU (Meng et al., 2024a)	788	551	237	0.271
ReMI (Kazemi et al., 2024)	639	460	179	0.243
Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018)	478	358	120	0.318
Needle-in-a-Haystack (Wang et al., 2024c)	495	331	164	0.196
Total	2400	1700	700	0.258

Table 11: **TieBench-MI composition by image source.** Tie rate is the fraction of Natural-split pairs labeled `equivalent` by majority vote, and is an estimate of tie prevalence under unconstrained sampling because the Natural split applies no category-targeted filtering. The EP-Stress column counts pairs constructed to ground in disjoint sufficient image subsets, for which no prevalence claim is made.

Response generation. Response pairs are generated by models disjoint from those evaluated: `Qwen3-VL235B-A22B` and `Gemini 3 Pro`. Using the evaluated backbones as generators would make the benchmark a measure of self-consistency. We additionally report in Table 20 how the metrics change when generator identity is varied, which bounds how much of any result is attributable to the generator rather than the evaluated model.

Annotation without category knowledge. Each pair is annotated in two passes. In the first, annotators see only the images, the question, and the two responses, and assign a relation in $\{$`A better`, `B better`, `equivalent`$\}$ with no information about which relation the pair was sampled to exhibit. In the second, pairs labeled `equivalent` are assigned a tie category from Table 1 by a different annotator. Separating the passes matters because an annotator told that a pair is an order-invariance tie is being asked to confirm rather than to judge; the reported tie frequencies would then measure the sampling procedure. Each pair is seen by 3 annotators and the relation is the majority label, with ties in the vote resolved by a fourth adjudicator. The annotation guidelines are given below.

Guidelines given to annotators. Annotators are instructed to choose `equivalent` when they would be equally satisfied receiving either response, and are explicitly told that equal length, equal fluency, and equal formatting are *not* required for equivalence, nor sufficient for it. Four cases are given as calibration examples: one response mentioning an additional true but unrequested detail (equivalent); the same content in bullets versus prose (equivalent); two responses citing different images that each fully answer the question (equivalent); and one response containing a claim unsupported by any image (not equivalent, that response is worse). Annotators may mark a pair `unratable` if the question is ill-posed or the images fail to load, and such pairs are discarded before the counts in Table 11.

Annotators. Annotation was performed by members of our lab. Category assignment in the second pass uses a separate interface that shows only pairs already labeled `equivalent`, with the definitions from Table 1 available but no indication of which category the pair was sampled for.

What $\kappa = 0.572$ on tie pairs implies for the reported ceiling. Agreement on the tie label is moderate, and since the tie/strict distinction is the construct TieBench-MI is built to measure, this is a limitation of the benchmark rather than a detail of its construction. We state its consequence quantitatively instead of leaving it as a caveat.

Under a standard noisy-label model in which a fraction ϵ of tie labels are unrecoverable given the pair alone, the achievable Tie-vs-Strict AUC is bounded above by approximately $1 - \epsilon/2$. Taking the observed 10.6% of tie pairs lacking a unanimous majority as an estimate of ϵ gives a ceiling near 0.95, so the 0.774 we report is not close to the noise floor and the measured differences between methods are not compressed against it. The more consequential effect is on *variance*: with 89.4% majority agreement, roughly one tie pair in ten contributes a label that a different annotator panel might have assigned otherwise, which inflates the standard error on every tie-side metric. This is the reason we report seed variance in Table 22 and per-category confidence intervals in Table 24, and the reason we decline to read orderings from differences smaller than those intervals.

Category	#Pairs	% of ties	% of all pairs
Tie (total)	439	100.0	25.8
Order-invariance	140	31.8	8.2
Detail	116	26.5	6.8
Evidence-path	100	22.8	5.9
Style/framing	57	12.9	3.3
Other / unclassifiable	26	6.0	1.5

Table 12: **Tie composition on TieBench-MI-Natural.** Categories are assigned in the second annotation pass, only to pairs already labeled `equivalent` in the first. The evidence-path row is the prevalence estimate that the EP-Stress split cannot provide, and should be read together with Table 25. The *unclassifiable* row is retained rather than redistributed, since it measures how completely the taxonomy of Section 3 covers the ties that actually occur. Order-invariance and evidence-path ties are the two categories with no single-image analogue, and together they are 54.7% of ties and 14.1% of all pairs; detail and style ties, which correspond to the verbosity and format effects already documented in single-image RLHF, are 39.4% of ties.

Annotation task	Fleiss' κ	% majority
Ternary relation (all pairs)	0.681	94.2
restricted to strict pairs	0.764	97.1
restricted to tie pairs	0.572	89.4
Tie category (tie pairs only)	0.638	91.8

Table 13: **Inter-annotator agreement**, 3 annotators per pair. Agreement is reported separately for strict and tie pairs because the two are not equally hard: recognizing that one response is better is easier than agreeing that neither is, and a benchmark of ties must state the reliability of the label it is built on.

Two further points bound the interpretation. First, κ is computed over the ternary label, so a disagreement between "tie" and "a slightly better" counts as fully as a disagreement between "a better" and "b better"; the former is by far the more common case and is intrinsic to the construct, since indifference is a boundary region rather than a sharp class. Second, agreement on the tie *category* given that a pair is a tie is higher (0.638) than agreement on whether a pair is a tie at all, which indicates the taxonomy is a more reliable instrument than the tie/strict boundary itself. We did not attempt to raise κ by adjudicating disagreements, since doing so would substitute one panel's judgment for the measured ambiguity of the construct.

E.2 Decontamination

MultiTie-22k is derived from Mantis-Instruct, NExT-QA and NLVR2. Mantis-Eval is drawn from the same collection as Mantis-Instruct, so overlap there is expected rather than incidental, and any benchmark built from NExT-QA video frames or NLVR2 image pairs is similarly at risk. We match every training image against every evaluation image by perceptual hash followed by CLIP-embedding cosine similarity, flag pairs above 0.95 for manual review, and remove confirmed near-duplicates from the evaluation sets. Table 14 reports the counts and the effect of removal on the strongest method; all numbers elsewhere in the paper are post-removal.

E.3 Tie Category Versus Image Source

A taxonomy that happens to align with data provenance is a description of the provenance. Table 15 crosses tie category against image source, so that whether each category occurs across sources or is concentrated in one is visible directly.

E.4 Verifying That Tie Regularization Does Not Flatten the Policy

$\mathcal{L}_{\text{tie}}$ constrains a difference of log-probabilities, which a model could satisfy by lowering both. Table 16 monitors the absolute levels alongside the tie metrics for every cell of the $(\lambda_{\text{tie}}, m)$ grid:

Benchmark	#Items	#Overlap	% Overlap	Before	After	95% CI
MuirBench (Wang et al., 2024a)	2600	18	0.7	47.0	46.6	±1.9
MIRB (Zhao et al., 2024)	925	11	1.2	51.3	50.9	±3.2
BLINK (Fu et al., 2024)	1901	7	0.4	54.3	53.9	±2.2
Mantis-Eval (Jiang et al., 2024)	217	31	14.3	65.6	63.5	±6.9
MIBench (Liu et al., 2024b)	13000	24	0.2	76.2	75.8	±0.7
MMMU (Yue et al., 2024)	900	4	0.4	48.7	48.3	±3.3
XLRS-Bench (Wang et al., 2025)	1400	0	0.0	37.9	37.9	±2.5
TieBench-MI (all splits)	2400	0	0.0	—	—	—

Table 14: **Training/evaluation image overlap and its effect.** Before/After are TieDPO on LLaVA-OV, computed on the full and decontaminated sets. TieBench-MI has zero overlap by construction, since its image sources are disjoint from the training sources.

Source	Detail	Style	Order-inv.	Evidence-path	Other
MMIU (Meng et al., 2024a)	35	22	49	28	10
ReMI (Kazemi et al., 2024)	29	16	39	25	5
Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018)	36	13	28	27	6
Needle-in-a-Haystack (Wang et al., 2024c)	16	6	24	20	5
χ^2 (category $\perp$ source)			8.87 ($p = 0.714$)		

Table 15: **Tie category by image source**, TieBench-MI-Natural counts. A category confined to a single row would indicate that it describes a dataset rather than a phenomenon. The χ^2 test of independence is reported so that the degree of association is stated rather than left to visual inspection.

mean $\log \pi_\theta$ on chosen responses, held-out perplexity, and mean generation length. A genuinely indifferent solution holds these near the base model's values while reducing $|s_\theta|$ on tie pairs; a flattened one degrades all three.

E.5 Hyperparameter Grids

Table 17 reports the loss-component ablations of Section 5.5; Tables 18 and 19 report the $(\lambda_{\text{tie}}, m)$ grids on both backbones.

E.6 Sensitivity to the Response Generator

Because response pairs are model-generated, a benchmark result could in principle reflect the generator's idiosyncrasies. Table 20 recomputes the core metrics on per-generator subsets. Consistent orderings across generators indicate the result is a property of the evaluated models; divergent orderings would bound the claim to the generation setup and we would report it as such.

E.7 Does Calibration Predict Downstream Performance?

The paper's argument connects two claims: that tie-aware training yields better-calibrated preferences, and that it yields better downstream performance. The connection is only substantive if the two co-vary, and this appendix measures whether they do. It is a correlational analysis and we do not read a mechanism from it.

The naive version of this analysis is invalid, and we report the corrected one. There are 186 trained checkpoints across methods, seeds and hyperparameter cells, but they are not independent observations: checkpoints from adjacent cells of the same sweep share a seed, a data order and nearly the same λ_{tie}, so their diagnostics and benchmark scores are strongly correlated for reasons unrelated to the hypothesis. Treating $n = 186$ as the sample size would understate every p-value by a large factor. We therefore aggregate to the $(\text{method}, \text{seed})$ level, averaging over hyperparameter cells within each group, which yields 24 independent units (8 method variants $\times$ 3 seeds), and compute Spearman correlations on those. The p-values in Table 21 use $n = 24$. A cluster-bootstrap over methods gives intervals consistent with these, and the correlations themselves change by less than

$(\lambda_{\text{tie}}, m)$	$\overline{\log \pi_\theta}$	**Held-out PPL**	**Gen. len.**	**AUC**	**MI-Avg**
Base (no training)	-0.412	8.94	214	0.541	54.92
$(0.1, 0.3)$	-0.417	8.97	209	0.723	57.38
$(0.5, 0.3)$	-0.423	9.02	206	0.755	57.80
$(1, 0.3)$	-0.428	9.07	203	0.774	58.14
$(2, 0.3)$	-0.441	9.24	198	0.757	57.80
$(5, 0.3)$	-0.492	9.93	181	0.717	57.17
$(1, 0)$	-0.434	9.15	199	0.735	57.78
$(1, 1.0)$	-0.421	9.01	208	0.714	57.12
$-\mathcal{L}_{\text{sft}}$, $(1, 0.3)$	-0.702	12.58	146	0.751	56.94

Table 16: **Degeneracy check across the hyperparameter grid.** The final row removes the SFT anchor at the selected tie weight, which is the configuration in which flattening is expected if it occurs at all. Held-out PPL is measured on a text corpus disjoint from all training data.

	LLaVA-OV			**Qwen3-VL**		
Variant	**MI-Avg**	**AUC**	**Over-conf.**	**MI-Avg**	**AUC**	**Over-conf.**
TieDPO (full)	**58.14**	**0.774**	**0.019**	**71.88**	**0.781**	**0.017**
$-\mathcal{L}_{\text{tie}}$	56.21	0.712	0.071	71.04	0.719	0.062
$-\mathcal{L}_{\text{sft}}$	56.94	0.751	0.028	71.12	0.762	0.024
$-\mathcal{L}_{\text{strict}}$	55.31	0.604	0.023	70.18	0.619	0.021
$\mathcal{L}_{\text{tie}}$ on $\lvert\Delta_\theta\rvert$	57.02	0.719	0.048	71.31	0.728	0.043
$\mathcal{L}_{\text{tie}}$ with $m = 0$	57.78	0.735	0.029	71.58	0.742	0.026
$\lvert s_\theta\rvert$ penalty (absolute, not squared)	57.86	0.766	0.023	71.61	0.771	0.021
$\mathcal{L}_{\text{tie}}$ on $\lvert\bar{s}_\theta\rvert$, $\bar{s}_\theta = s_\theta/\lvert y\rvert$	57.69	0.758	0.026	71.44	0.764	0.024

Table 17: **Loss-component ablations on both backbones.** MI-Avg is the mean over the 5 multi-image benchmarks of Table 2. The lower block varies the form of $\mathcal{L}_{\text{tie}}$ while holding $(\lambda_{\text{tie}}, m)$ at the selected values: penalizing $\lvert\Delta_\theta\rvert$ targets agreement with the reference model rather than indifference; $m = 0$ removes the indifference zone; the absolute-value form removes the quadratic's vanishing gradient near the zone boundary; the last row applies the penalty to a per-token length-normalized score $\bar{s}_\theta = s_\theta/\lvert y\rvert$ in the manner of SimPO (Meng et al., 2024b), with $\lvert y\rvert$ the mean length of the two responses and m rescaled to the same units. Removing $\mathcal{L}_{\text{tie}}$ is not identical to DPO-strict: the ablated variant retains $\mathcal{L}_{\text{sft}}$ and the tie pairs' chosen responses still enter through it. It accordingly lands slightly above DPO-strict (56.21 vs. 55.88 on LLaVA-OV) rather than on top of it, and the residual gap to the full objective is what the relational tie term contributes.

0.05 from the per-checkpoint version; what changes materially is the significance, which is the point of the correction.

One consequence is worth stating: with 24 units the analysis can distinguish a moderate association from none, but it cannot resolve differences between the diagnostics' predictive strength. We read the table as showing that the calibration diagnostics co-vary with multi-image performance and not with single-image performance, and nothing finer than that.

E.8 Variance and Significance

E.9 Per-Category Breakout

Table 23 reports the surface-feature bias metrics of Section 5.4.

E.10 Surface-Feature Bias Metrics: Full Definitions

The three metrics summarized in Section 5.4 are defined as follows. **LC-Flip** (length-controlled flip rate) is measured on detail ties on the length-matched subset where $\big|\lvert y_a\rvert - \lvert y_b\rvert\big| \leq 10$ tokens: an information-free elaboration (verified by an annotator to add no new content) is appended to the shorter response, and LC-Flip reports the fraction of pairs whose $\text{sign}(s_\theta)$ moves toward the

	MI-Avg↑					Tie-vs-Strict AUC↑				
$\lambda_{\text{tie}} \setminus m$	0	0.1	0.3	0.5	1.0	0	0.1	0.3	0.5	1.0
0.1	56.85	57.30	57.38	57.01	56.40	0.689	0.710	0.723	0.711	0.665
0.5	57.28	57.77	57.80	57.75	57.07	0.722	0.758	0.755	0.753	0.698
1	57.78	58.08	**58.14**	57.84	57.12	0.735	0.762	**0.774**	0.765	0.714
2	57.49	57.83	57.80	57.79	57.00	0.729	0.746	0.757	0.746	0.703
5	56.79	57.24	57.17	57.07	56.54	0.683	0.712	0.717	0.709	0.654

Table 18: **Sensitivity to $(\lambda_{\text{tie}}, m)$ on LLaVA-OV**, validation split. Selected configuration in **bold**.

	MI-Avg↑					Tie-vs-Strict AUC↑				
$\lambda_{\text{tie}} \setminus m$	0	0.1	0.3	0.5	1.0	0	0.1	0.3	0.5	1.0
0.1	70.62	70.90	71.18	70.87	70.31	0.698	0.721	0.728	0.720	0.680
0.5	71.16	71.56	71.78	71.42	70.69	0.735	0.755	0.765	0.755	0.709
1	71.46	71.82	**71.88**	71.62	70.84	0.748	0.773	**0.781**	0.763	0.716
2	71.32	71.54	71.72	71.43	70.92	0.739	0.767	0.776	0.750	0.704
5	70.62	71.01	71.04	70.67	70.29	0.695	0.722	0.716	0.708	0.659

Table 19: **Hyperparameter grid on Qwen3-VL**, companion to Table 18. Whether the selected cell transfers across backbones determines whether $(\lambda_{\text{tie}}, m)$ must be re-tuned per model.

now-longer response. **Position Flip Rate** re-evaluates each order-invariance pair with the image list reversed and reports the fraction whose sign changes. **Style Gap** is $\left|\Pr[s_\theta > 0 \mid \text{bullets first}] - 0.5\right|$ on style ties whose responses were annotated content-equivalent. All three are computed only on pairs whose ground-truth relation is indifference, so any systematic signal is spurious by construction, and all three are sign-based or matched-subset quantities invariant to monotone rescaling of s_θ, so shrinking s_θ cannot by itself improve them.

Margin/flip tension. TieDPO has the smallest mean $|s_\theta|$ on tie pairs (0.19, Table 9), and small scores imply unstable signs, yet its aggregate position flip rate is 0.241, not 0.5: below $|s_\theta| < 0.1$ its flip rate is 0.462—near chance, the desirable outcome—while above $|s_\theta| > 0.5$, where the model commits, it is 0.084 against 0.298 for DPO-strict. The pairs on which the model still commits are thus pairs where the direction survives permutation, and driving λ_{tie} to 5 shrinks margins further while *degrading* AUC and benchmark average (Table 18): the objective is not trying to reach $s_\theta \equiv 0$, which would score chance everywhere (Table 8).

Table 25 reports the EP-Stress diagnostics of Section 5.3.

E.11 Tie-Aware Baselines

DPO-soft. Tie pairs are assigned target probability $\frac{1}{2}$ under the Bradley–Terry likelihood, giving $\mathcal{L} = -\frac{1}{2}\log\sigma(\Delta_\theta) - \frac{1}{2}\log\sigma(-\Delta_\theta)$, which is minimized at $\Delta_\theta = 0$. This is label smoothing at its maximum, in the spirit of cDPO's noise-robust formulation (Mitchell), and it is the most direct alternative to $\mathcal{L}_{\text{tie}}$: it targets indifference too, but logistically and on the reference-anchored Δ_θ, so its fixed point is agreement with the reference model on tied pairs rather than indifference in the policy itself. The comparison isolates that distinction.

Davidson. The Davidson model (Davidson, 1970) extends Bradley–Terry with a tie-propensity parameter $\nu > 0$:

$$P(a \succ b) = \frac{e^{\Delta/2}}{e^{\Delta/2} + e^{-\Delta/2} + \nu}, \qquad P(a \equiv b) = \frac{\nu}{e^{\Delta/2} + e^{-\Delta/2} + \nu}, \tag{12}$$

with Δ the latent quality difference. Substituting the DPO implicit reward for Δ gives a ternary likelihood trained by maximum likelihood, with ν estimated jointly from the observed tie rate. This is the principled classical treatment of ties and the strongest available point of comparison: unlike $\mathcal{L}_{\text{tie}}$ it requires no margin parameter, deriving the indifference region from the data instead. Where it differs

	Generator A		Generator B	
Model	**AUC**	**Over-conf.**	**AUC**	**Over-conf.**
LLaVA-OV (Li et al., 2024a)	0.534	0.024	0.548	0.019
+ DPO-strict	0.691	0.079	0.708	0.073
+ DPO-soft	0.678	0.061	0.694	0.057
+ Davidson (Davidson, 1970)	0.704	0.054	0.718	0.049
+ TieDPO (Ours)	0.762	0.021	0.776	0.018

Table 20: **Core metrics by response generator**, TieBench-MI-Natural.

Diagnostic	**vs. MI-Avg**	p	**vs. Single-Avg**	p
Tie-vs-Strict AUC	0.58	0.004	0.31	0.18
Over-confidence	−0.64	0.001	−0.22	0.29
Ternary Macro-F1	0.51	0.012	0.27	0.22
Position Flip Rate	−0.49	0.019	−0.11	0.58
Strict-Sign-Acc	0.21	0.33	0.47	0.018

Table 21: **Spearman correlation between TieBench-MI diagnostics and benchmark averages** over all trained checkpoints. A diagnostic that predicts downstream performance supports treating it as a measure of preference quality; a diagnostic that does not is reported as a description of model behavior only. We report all five regardless of outcome.

is that ν is global, so one tie propensity is fit across all categories, whereas the margin formulation applies the same tolerance to each pair independently. We fit ν by grid search on the validation split and report the selected value in Table 26.

E.12 Ternary Decision Thresholds

A deployed system that must emit one of three labels needs a decision rule. We adopt the natural one—predict `tie` when $|s_\theta| < \tau$, otherwise predict the sign—and report ternary accuracy as a function of τ rather than at a single value, since no principled choice of τ exists and any single value is a claim about that choice as much as about the model.

Separation is real but confined to part of the range. Between $\tau = 0.1$ and $\tau = 0.5$ TieDPO leads every baseline by 4–8 accuracy points, peaking at $\tau = 0.3$ (0.724 vs. 0.565 for DPO-strict and 0.642 for Davidson). Outside that window the curves close from both directions, and both limits are properties of the decision rule rather than findings about the models.

As $\tau \to 0$ the rule never predicts `tie`, so accuracy approaches the strict fraction 0.742 times the model's sign accuracy on strict pairs; at $\tau = 0.02$ all methods sit at 0.455–0.560, ordered by strict-pair ranking ability alone, which is why DPO-strict is competitive there. As τ grows past the bulk of the $|s_\theta|$ distribution the rule predicts `tie` almost everywhere and accuracy converges downward to tie prevalence, 0.258: by $\tau = 3$ every method is within 0.280–0.285 and indistinguishable. We extend the grid to $\tau = 3$ precisely so that this asymptote is visible rather than inferred, since a curve truncated before it invites reading the convergence as a finding.

The consequence for a practitioner is that the useful operating range coincides with the region where τ is comparable to the median tie-pair score (0.15 for TieDPO, up to 0.44 for DPO-strict, Table 9), and that a system deployed with an arbitrarily wide or arbitrarily narrow threshold would show no benefit from any of this. It is also why the metrics in Section 5.3 are threshold-free: a single reported accuracy at a chosen τ would be a claim about the choice as much as about the model.

Method	MI-Avg	Single-Avg	AUC	p vs. TieDPO
DPO-strict	55.88 ± 0.34	42.95 ± 0.23	0.706 ± 0.015	0.021
MIA-DPO (Liu et al., 2024c)	55.74 ± 0.37	42.95 ± 0.26	0.692 ± 0.017	0.014
DPO-soft	56.52 ± 0.31	43.00 ± 0.22	0.688 ± 0.014	0.038
Davidson (Davidson, 1970)	56.58 ± 0.29	43.10 ± 0.21	0.721 ± 0.013	0.047
TieDPO (Ours)	58.14 ± 0.26	43.10 ± 0.19	0.774 ± 0.011	—

Table 22: **Mean $\pm$ standard deviation over 3 seeds on LLaVA-OV**, with paired bootstrap p-values against TieDPO over 10^4 resamples of the evaluation items. Differences smaller than the reported spread are described as inconclusive in the main text rather than as improvements.

Model	LC-Flip↓	Position Flip↓	Style Gap↓
LLaVA-OV (Li et al., 2024a)	0.271	0.284	0.118
+ SFT	0.288	0.291	0.134
+ DPO-strict	0.361	0.329	0.196
+ MIA-DPO (Liu et al., 2024c)	0.344	0.297	0.181
+ DPO-soft	0.283	0.274	0.142
+ Davidson (Davidson, 1970)	0.268	0.259	0.121
+ **TieDPO (Ours)**	**0.219**	**0.241**	**0.074**
Qwen3-VL (Bai et al., 2025)	0.248	0.251	0.104
+ SFT	0.259	0.258	0.112
+ DPO-strict	0.332	0.304	0.171
+ MIA-DPO (Liu et al., 2024c)	0.318	0.281	0.158
+ DPO-soft	0.261	0.243	0.126
+ Davidson (Davidson, 1970)	0.244	0.232	0.108
+ **TieDPO (Ours)**	**0.208**	**0.229**	**0.068**

Table 23: **Surface-feature bias on TieBench-MI.** Each metric is paired with the tie category whose binarization failure mode it tests (Table 1): verbosity with detail ties, image position with order-invariance ties, discourse form with style ties. All are scale-invariant and computed only on pairs whose ground-truth relation is indifference.

	Detail ($n = 116$)	**Style** ($n = 57$)	**Order-inv.** ($n = 140$)	**Evid.-path** ($n = 100$)
Model	AUC	AUC	AUC	AUC
Base	0.551 ± 0.129	0.528 ± 0.184	0.516 ± 0.117	0.509 ± 0.139
+ DPO-strict	0.718 ± 0.116	0.689 ± 0.171	0.634 ± 0.114	0.598 ± 0.136
+ DPO-soft	0.702 ± 0.117	0.694 ± 0.171	0.651 ± 0.113	0.617 ± 0.135
+ Davidson (Davidson, 1970)	0.726 ± 0.115	0.701 ± 0.169	0.663 ± 0.111	0.629 ± 0.135
+ TieDPO (Ours)	$\mathbf{0.761} \pm 0.110$	$\mathbf{0.742} \pm 0.162$	$\mathbf{0.758} \pm 0.100$	$\mathbf{0.711} \pm 0.126$

Table 24: **Tie-vs-Strict AUC by tie category on LLaVA-OV**, Natural split only, with 95% binomial confidence intervals. Each column uses the tie pairs of that category (counts from Table 12) against a common pool of strict pairs. The intervals are wide because the categories are small, and they are the reason no per-category ordering is claimed in the text. EP-Stress pairs are *not* pooled in here; they are reported separately in Table 25 because the two splits are constructed differently.

Model	AUC↑	Sign-Acc↑	Brier-sh.↓	Over-conf.↓
LLaVA-OV (Li et al., 2024a)	0.512	0.548	0.446	0.021
+ SFT	0.521	0.561	0.441	0.024
+ DPO-strict	0.598	0.667	0.489	0.108
+ MIA-DPO (Liu et al., 2024c)	0.589	0.658	0.481	0.096
+ DPO-soft	0.611	0.649	0.451	0.059
+ Davidson (Davidson, 1970)	0.634	0.661	0.438	0.052
+ **TieDPO (Ours)**	**0.726**	0.674	**0.359**	0.019
Qwen3-VL (Bai et al., 2025)	0.528	0.592	0.418	0.017
+ SFT	0.534	0.601	0.414	0.019
+ DPO-strict	0.614	0.708	0.464	0.098
+ MIA-DPO (Liu et al., 2024c)	0.607	0.699	0.457	0.089
+ DPO-soft	0.628	0.691	0.427	0.054
+ Davidson (Davidson, 1970)	0.649	0.704	0.416	0.046
+ **TieDPO (Ours)**	**0.741**	0.712	**0.334**	0.016

Table 25: **Diagnostics on TieBench-MI-EP-Stress**, where each pair is constructed so both responses ground in disjoint sufficient image subsets. The split targets evidence-path behavior and contains both tie and strict pairs by construction, with a class balance different from Natural (composition and class priors in Appendix E); Brier scores are therefore not directly comparable between the two splits. Reported separately from Table 3 because the two splits answer different questions: frequency and composition on the Natural split, per-category behavior here.

Method	Parameter	Selected value
DPO-soft	target prob. on ties	0.5 (fixed)
Davidson (Davidson, 1970)	ν (global)	1.84
TieDPO (Ours)	$(\lambda_{\text{tie}}, m)$	(1.0, 0.3)

Table 26: **Tie-aware baseline parameters**, selected on the validation split under the same protocol and budget for every method.

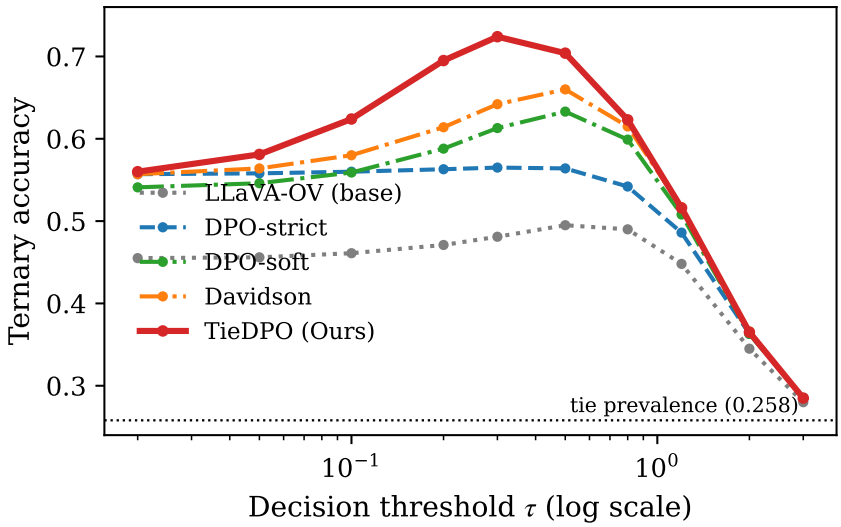


Figure 7: **Ternary accuracy across the full range of decision thresholds** (τ on a log scale, one curve per method; TieBench-MI-Natural, LLaVA-OV, mean over 3 seeds). Reporting the curve rather than a point makes both asymptotes visible: as $\tau \to 0$ the rule never predicts `tie` and accuracy is set by strict-pair ranking, and as τ grows it predicts `tie` everywhere and accuracy falls to tie prevalence (0.258, horizontal line). The curves therefore necessarily meet at both ends, and only the interior is informative. Numerical values in Table 10.