BEYOND BINARY PREFERENCES: GRADED PREFERENCE OPTIMIZATION FOR LIMB-MOTION CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) can generate rich video captions, yet often misidentify which person performs an action or which limb is involved, particularly across camera cuts. Improving these details requires evaluation and training that distinguish missing information from incorrect assertions. We introduce FlexBench, a benchmark spanning 3,105 shots and 18,161 evaluation queries, with human-verified identities and systematic per-person coverage of fine-grained limb actions and states. Its reference-derived checklists support automated assessment of complete captions in their person and shot contexts. Our Graded Physical Alignment score (GPA) awards credit for correct content and deducts points for incorrect or fabricated actions, making these errors explicit in the aggregate score. Building on this rubric, we propose Graded Margin Direct Preference Optimization (GM-DPO), which assigns stronger preference margins and greater training weight to more severe action errors. Across three VLM backbones, GM-DPO achieves the highest substantive-action and GPA scores among the evaluated preference objectives, improving GPA over DPO by 2.02–3.40 points. On Qwen3-8B, it reduces the weighted hallucination rate by 21.3% relative to DPO. These gains accompany sustained long-form output, improved shot structure, and competitive performance on three additional multimodal benchmarks.

1 INTRODUCTION

Video captioning turns visual observations into language, making video content accessible for understanding and downstream learning. Recent Vision-Language Models (VLMs) can produce rich captions covering scenes, people, and events (Team, 2025; Clark et al., 2026), yet describing how an action unfolds remains challenging. Recognizing that someone opens a box, for example, does not establish which hand lifts the lid or how the other hand supports it. Across camera cuts and interacting people, these details must also remain attached to the correct person and shot. Our goal is to improve limb-motion fidelity within dense captions generated by general-purpose VLMs.

Progress toward this goal requires both targeted supervision and suitable evaluation. Supervised fine-tuning (SFT) teaches caption generation through reference imitation; subsequent preference optimization can further improve fidelity by contrasting better and worse captions (Yuan et al., 2025; Lee et al., 2025). Direct Preference Optimization (DPO) (Rafailov et al., 2023) makes this practical with fixed preference pairs, without a separate reward model or online sampling during optimization. However, its binary preference labels do not explicitly distinguish omissions, limb confusions, and fabricated actions. This distinction matters when captions become training annotations: missing information leaves supervision incomplete, whereas false assertions introduce incorrect video–text associations. Evaluation presents a related challenge. Widely used multiple-choice question answering (MCQA) benchmarks (Li et al., 2024; Hong et al., 2025) measure answer selection, which does not directly reveal what a model would assert in an unrestricted caption. Although recent benchmarks directly evaluate motion captions (Tu et al., 2025; Lin et al., 2026), systematic limb-level assessment of every identifiable person across shots remains insufficiently addressed.

As shown in Figure 1, we address these needs through complementary advances in evaluation and training. We introduce FlexBench, a Fine-grained Limb-motion EXamination Benchmark, with human-verified identities and reference-derived checklists covering each person's actions in their corresponding shots. Coverage extends to stationary limbs and limb visibility, ensuring that evalua-

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.

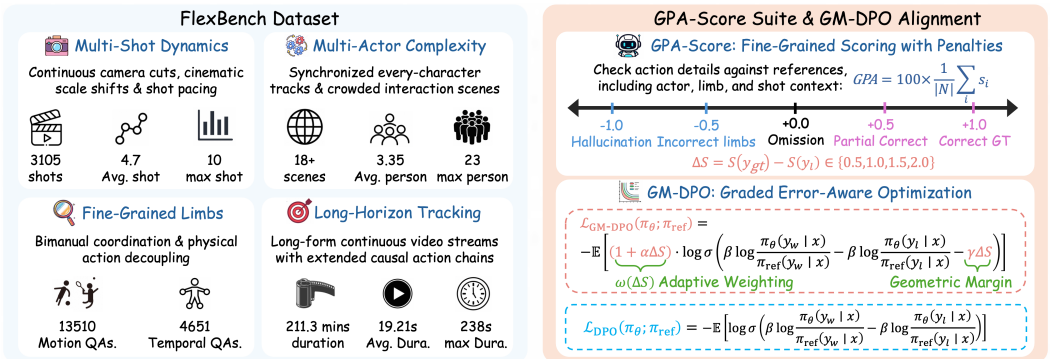


Figure 1: Overview of our framework. FlexBench evaluates fine-grained limb motion for individual people across shots. GPA aggregates graded factual credit and explicit error penalties, while GM-DPO uses action-error severity to adjust preference margins and loss weights.

tion encompasses limb states as well as movements. Our GPA combines fine-grained grades through a weighted average, awarding full or partial credit, assigning zero credit to omissions, and deducting points for incorrect or fabricated actions. Guided by the same grading rubric, GM-DPO assigns larger preference margins and loss weights to more severe action errors. This focuses preference learning on limb-motion fidelity within complete, richly detailed video captions.

Our main contributions are three-fold:

- **FlexBench:** A multi-shot benchmark spanning 3,105 shots and 18,161 evaluation queries, with systematic per-person coverage of fine-grained limb motion, human-verified cross-shot identities, and checklists addressing actions, stationary limbs, and visibility.
- **GPA:** A graded, penalty-aware metric for automatically evaluating generated captions against detailed references, distinguishing incomplete coverage from incorrect motion claims.
- **GM-DPO:** An offline preference objective that incorporates action-error severity into margins and loss weights. Across three backbones, it achieves the best substantive-action and GPA scores among tested preference objectives while sustaining long-form output.

2 Related Work

2.1 From Video Understanding to Limb-Motion Captioning

General-purpose VLMs increasingly capture not only video events but also their temporal structure, with benchmarks assessing long-video comprehension (Fu et al., 2025; Wu et al., 2024) and motion perception (Hong et al., 2025). Limb-motion captioning requires finer granularity: decomposing activities into constituent actions and grounding each action in the correct person, limb, and interaction. KPM-Bench (Lin et al., 2026) advances fine-grained motion captioning, accompanied by a training framework combining multi-level motion representations with SFT and GRPO. Our work addresses the complementary challenge of maintaining limb-level accuracy within dense captions that cover every identifiable person across shots. This scope includes stationary limbs and visibility, enabling assessment of both action coverage and unsupported motion claims.

2.2 Preference Optimization for Video Captioning

Preference learning complements reference imitation with explicit comparisons between candidate outputs. DPO (Rafailov et al., 2023) provides an offline objective for this supervision, while c-DPO, IPO, and SimPO explore label smoothing, squared preference objectives, and length-normalized rewards, respectively (Mitchell, 2023; Azar et al., 2024; Meng et al., 2024). For video captioning, Tarsier2 (Yuan et al., 2025) applies DPO after supervised training to improve detailed captions. VidChain (Lee et al., 2025) combines supervised captioning and temporal grounding with metric-based DPO, using task metrics to select preference pairs. VideoComp (Kim et al., 2025) instead trains video–text matching models with a hierarchical pairwise preference loss over increasingly disrupted captions. These studies motivate targeted preference supervision for video understanding. GM-DPO uses a limb-motion error rubric to determine both the preference margin and the loss weight, distinguishing incomplete action content from incorrect or fabricated actions.

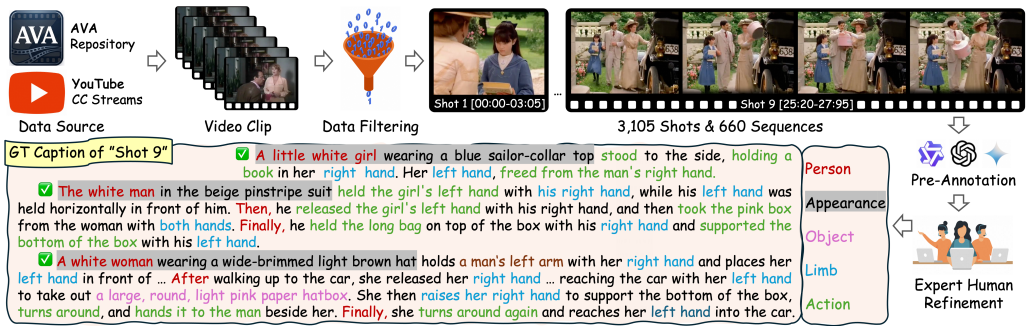


Figure 2: FlexBench curation pipeline: video filtering, shot boundary calibration, VLM pre-labeling, and multi-stage expert refinement for granular limb-action modeling.

2.3 Motion Benchmarks and Caption Evaluation

General video benchmarks assess broad perceptual and temporal capabilities (Li et al., 2024; Fu et al., 2025; Wu et al., 2024). Motion-focused evaluations examine fine-grained dynamics via multiple-choice queries (MotionBench (Hong et al., 2025)) and diverse QA-caption tasks (TempCompass (Liu et al., 2024)). Free-form generation additionally exposes the action claims a model produces without candidate answers. FAVOR-Bench (Tu et al., 2025) evaluates motion captions through LLM-assisted scoring and a separate structured sequence-matching protocol. Bridging these gaps, FlexBench emphasizes comprehensive, per-person and shot-specific limb assessments (covering stationary and hidden states). Our GPA metric resolves error severity via graded credits and penalties, supported by a temporal suite for consecutive-action coverage.

3 FlexBench: Fine-Grained Limb Physical Benchmarking

FlexBench evaluates fine-grained limb-motion fidelity within complete multi-shot video captions. By pairing verified references with detailed checklists, it audits each person's active limbs and exact shot locations—enabling rigorous motion assessment without sacrificing descriptive richness.

3.1 Data Collection and Quality Assurance Pipeline

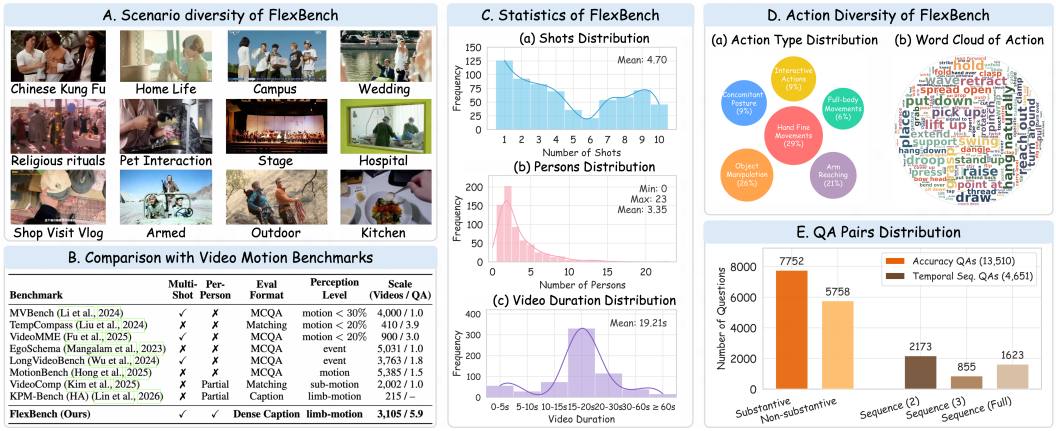
To prevent models from exploiting static background shortcuts, video sequences in FlexBench are curated from untrimmed cinematic scenes in AVA (Gu et al., 2018) and CC-licensed YouTube streams, featuring domestic tasks, crafts, and complex tool manipulation. As shown in Figure 3A, the collection spans varied interactions. Automated shot detection with manual boundary calibration yields 660 sequences comprising 3,105 shots, with a mean sequence duration of 19.21 seconds.

As shown in Figure 2, 18 trained annotators refine VLM-generated captions to specify actor identities, interacted objects, limb laterality, and action order. Cross-shot identities are manually verified, with distinctive appearance cues used to distinguish similar-looking people. Annotations undergo blind review and consensus arbitration for ambiguous cases; details appear in Appendix A.

3.2 Per-Person Coverage and Checklist Evaluation

Coverage across people and shots. Unlike conventional suites limited to isolated cuts or salient actors, FlexBench pairs multi-shot continuity with systematic, per-person limb annotation (Figure 3B). As shown in Figure 3C, sequences span 1–10 shots (mean 4.70) and 0–23 actors (mean 3.35). Reference captions structure each individual's actions by shot—resolving granular limb movements and object interactions while accounting for stationary or hidden limb states.

Reference-derived factual checks. Human-verified captions are decomposed into positive checklist items covering actions, limb states, object interactions, and temporal order, each tied to a specific person and shot. During evaluation, an LLM judge scores the VLM-generated caption against both the checklist and ground-truth reference. Full credit strictly requires joint alignment across the action, person, and shot context—mismatched attributions receive no credit, while hallucinations and incorrect claims incur graded penalties. Human-evaluator agreement is analyzed in Section 5.4.



Benchmark	Multi-Shot	Per-Person	Eval Format	Perception Level	Scale (Videos / QA)
MVBench (Li et al., 2024)	✓	✗	MCQA	motion < 30%	4,000 / 1.0
TempCompass (Liu et al., 2024)	✗	✗	Matching	motion < 20%	410 / 3.9
VideoMME (Fu et al., 2025)	✓	✗	MCQA	motion < 20%	900 / 3.0
EgoSchema (Mangalam et al., 2023)	✗	✗	MCQA	event	5,031 / 1.0
LongVideoBench (Wu et al., 2024)	✓	✗	MCQA	event	3,763 / 1.8
MotionBench (Hong et al., 2025)	✗	✗	MCQA	motion	5,385 / 1.5
VideoComp (Kim et al., 2025)	✗	Partial	Matching	sub-motion	2,002 / 1.0
KPM-Bench (HA) (Lin et al., 2026)	✗	Partial	Caption	limb-motion	215 / –
FlexBench (Ours)	✓	✓	**Dense Caption**	**limb-motion**	**3,105 / 5.9**

Figure 3: Comprehensive overview of FlexBench: (A) Scenario coverage; (B) Dimensional comparison against existing video benchmarks; (C) Structural shot, actor, and duration distributions; (D) Kinematic action categories; (E) Action and temporal query counts.

3.3 Graded Evaluation

As shown in Figure 3E, FlexBench spans 18,161 queries. The Action Accuracy Suite comprises 7,752 substantive action queries and 5,758 non-substantive posture queries. The Temporal Sequence Suite contains 4,651 queries assessing consecutive actions in their shot contexts. Together, the two suites measure the fidelity of individual action claims and their joint coverage across sequences.

The Graded Physical Alignment score (GPA) assigns action-evaluation items grades $s_i \in \{-1, -0.5, 0, 0.5, 1\}$, distinguishing correct, partially specified actions, omissions, and contradictory or fabricated actions. For category $k \in \mathcal{K} = \{\text{sub}, \text{non}\}$, let N_k denote scored items we report

$$\mathcal{S}_{\text{GPA}} = 100 \times \sum_{k \in \mathcal{K}} w_k \cdot \left(\frac{1}{|N_k|} \sum_{i \in N_k} s_i \right) = 100 \times \left(0.6 \cdot \bar{S}_{\text{sub}} + 0.4 \cdot \bar{S}_{\text{non}} \right), \tag{1}$$

Thus, grades use the $[-1, 1]$ scale, whereas reported GPA scores lie in $[-100, 100]$. The clipped variant GPA_{clip} replaces s_i with $\max(0, s_i)$ in Equation 1, measuring credited content without negative penalties. We also report the weighted hallucination rate $\text{WHR} = \text{Ratio}_{-0.5} + 2\text{Ratio}_{-1}$, where Ratio_v is the proportion of scored items assigned grade v.

Temporal queries evaluate consecutive substantive actions within shot contexts. For example, a Seq_2 item earns credit when both constituent actions are correctly represented. We report $\text{TSA}_{\text{w}} = 0.2\text{Seq}_2 + 0.3\text{Seq}_3 + 0.5\text{Seq}_{\text{chain}}$, weighting two-action, three-action, and full-chain queries.

4 GM-DPO: Graded Margin Preference Optimization

In this section, we formulate GM-DPO, an error-aware offline alignment objective to suppress fine-grained physical hallucinations in multimodal video understanding. GM-DPO addresses the core pathology of prevailing preference alignment algorithms: uniform treatment of non-preferred trajectories, obscuring the severe consequences of bodily chirality reversals and motion confabulations.

4.1 Preliminaries and Limitations of Standard DPO

Given pairwise preferences $\mathcal{D} = \{(x, y_w, y_l)\}$ and a frozen reference policy π_{ref}, Rafailov et al. (2023) optimizes a parameterized policy $\pi_\theta(y \mid x)$ under the Bradley-Terry model by minimizing:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(h_\theta(x, y_w) - h_\theta(x, y_l) \right) \right], \tag{2}$$

where $h_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ denotes the implicit reward. Differentiating Equation 2 with respect to θ yields the parameter gradient:

$$\begin{cases} \nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta \cdot \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\sigma \left(h_\theta(x, y_l) - h_\theta(x, y_w) \right) \cdot \Delta_\theta \log \pi(y_w, y_l) \right], \\ \Delta_\theta \log \pi(y_w, y_l) \triangleq \nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x), \end{cases} \tag{3}$$

Crucially, standard DPO assumes uniform binary preferences ($y_w \succ y_l$), scaling updates solely by prediction error $\sigma(h_\theta(x, y_l) - h_\theta(x, y_w))$ regardless of defect severity. Consequently, it applies identical gradient penalties to benign omissions and fatal chirality inversions (e.g., swapping hands), failing to penalize critical physical violations without over-correcting harmless variations.

4.2 Perturbation Severity from the Grading Rubric

Each preference pair applies one error category e consistently to the targeted action descriptions within a selected shot, retaining the remaining caption. This restriction preserves the surrounding context and is intended to avoid preference pairs that are easily distinguished through errors distributed across multiple shots. The original targeted content has grade 1, and the perturbation is assigned $s_{\text{pert}}(e) \in \{0.5, 0, -0.5, -1\}$ under the action-grading rubric. We define

$$\Delta S = 1 - s_{\text{pert}}(e) \in \{0.5, 1, 1.5, 2\}. \tag{4}$$

This label represents local perturbation severity, not the GPA difference between complete captions. It uses the unscaled rubric, without the factor of 100 used for benchmark reporting. The preference objective below evaluates likelihoods over complete captions.

4.3 GM-DPO Objective and Optimization Dynamics

Rather than enforcing uniform margins, GM-DPO governs policy updates through two coordinated mechanisms modulated by ΔS: an internal dynamic geometric margin $\gamma \Delta S$ enforcing separation, and an external gradient weight $\omega(\Delta S) = 1 + \alpha \Delta S$ ($\alpha \geq 0, \gamma > 0$):

$$\mathcal{L}_{\text{GM-DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\omega(\Delta S) \log \sigma \left(h_\theta(x, y_w) - h_\theta(x, y_l) - \gamma \Delta S \right) \right]. \tag{5}$$

Differentiating Equation 5 with respect to model parameters θ yields the explicit gradient:

$$\nabla_\theta \mathcal{L}_{\text{GM-DPO}} = -\beta \cdot \mathbb{E}_{\mathcal{D}} \left[\underbrace{\omega(\Delta S)}_{\text{Scaling}} \cdot \underbrace{\sigma \left(h_\theta(x, y_l) - h_\theta(x, y_w) + \gamma \Delta S \right)}_{\text{Margin-Shifted Residual Error}} \cdot \Delta_\theta \log \pi(y_w, y_l) \right]. \tag{6}$$

Equation 6 reveals the dual modulation mechanics of GM-DPO for vision-language alignment:

(1) Margin-Shifted Residual Error: The offset $+\gamma \Delta S$ shifts the logistic saturation boundary. When $h_\theta(y_w) > h_\theta(y_l)$, standard DPO gradients rapidly vanish; in contrast, GM-DPO sustains gradients until clearing a severity-scaled margin, ensuring continuous optimization on subtle dynamics.

(2) Magnitude Scaling: The multiplier $\omega(\Delta S)$ linearly magnifies parameter updates for severe physical flaws (up to $(1 + 2\alpha)\times$ for confabulations), penalizing chirality flips more aggressively than harmless omissions. This error-sensitive scaling functions as an adaptive step size, guiding the policy away from catastrophic hallucination regimes without destabilizing language representation.

4.4 Preference Pair Construction

We construct 125k preference pairs from single- and multi-shot videos, using expert-verified captions as chosen responses. To prevent data contamination, the training pool is completely disjoint from FlexBench: it strictly excludes AVA footage and shares no source videos (e.g., films or YouTube streams) with the 660 evaluation sequences.

As illustrated in Figure 4, rule-based editing produces Laterality Neglect ($\Delta S = 0.5$) and Chirality Inversion (1.5). Schema-constrained Gemini-3.5-Flash rewriting generates Motion Omission (1), Subaction Drift (0.5), and Action Confabulation (2). Untargeted descriptions are retained, and rewrites are constrained to preserve the original style. A double-blind study yielded 49.60% human accuracy in distinguishing human-refined from LLM-modified captions; generation and verification protocols are provided in Appendix A.3.

5 Experiments

We evaluate whether GM-DPO improves limb-level action fidelity within long-form, shot-structured video captions. Our experiments first characterize existing models on FlexBench, then compare various preference optimization methods, examine the contributions of the proposed objective, and assess performance on broader multimodal benchmarks.

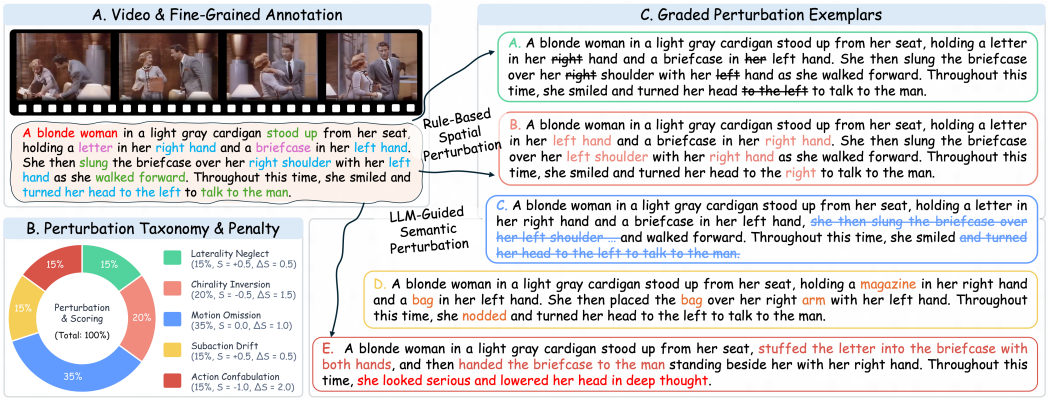


Figure 4: The preference pair curation pipeline and perturbation spectrum. Ground-truth dense captions ($y_w, S = +1.0$) are perturbed via rule-based spatial editing and schema-constrained LLM rewriting to synthesize negative samples across five graded physical error regimes ($\Delta S \in [0.5, 2.0]$).

5.1 Experimental Setup

Evaluated Model Suites. We benchmark physical motion perception across three representative foundation tiers: (1) *Proprietary Frontier APIs*: Gemini-3.1-Pro, Gemini-3.5-Flash (Google Gemini Team, 2026a;b), Seed-2.1-Pro (ByteDance Seed Team, 2026), Kimi-k2.6 (Kimi Team and Moonshot AI, 2026), alongside Qwen-3.5-Omni (Qwen Team, 2026a) and Qwen-3.8-Max (Qwen Team, 2026b); (2) *Open-Source Models*: Nemotron-3-Omni-30B (**?**), Gemma-4-31B (Team, 2026), Qwen3-32B (Team, 2025), and Qwen-3.8-27B (Qwen Team, 2026b). (3) *Open-Weight Video Models (7B)*: Tarsier2-Recap-7B (Yuan et al., 2025) and VideoLLaMA2-7B (Cheng et al., 2024). To ensure comparability and consistency, all models are evaluated under identical prompts.

Training Setup. We compare GM-DPO with DPO (Rafailov et al., 2023), c-DPO (Mitchell, 2023), IPO (Azar et al., 2024), and SimPO (Meng et al., 2024) on Qwen2.5-7B, Qwen3-8B, and LLaVA-2-8B, with an additional SFT baseline on Qwen3-8B. All preference methods include a chosen-response NLL term, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pref}} + \lambda \mathcal{L}_{\text{NLL}}$, with $\lambda = 1$. Experiments are conducted on a cluster of 8×NVIDIA A800 (80GB) GPUs using the AdamW optimizer, updating language backbone parameters while keeping visual encoders frozen. For both SFT and preference tuning, runs are repeated across three random seeds with mean metrics reported. Training configurations and baseline objectives appear in Appendix B; results with standard deviations are provided in Appendix C.1.

Evaluation Metrics. We report GPA, TSA_{w}, and their sub-metrics (averaged across GPT-4o, Gemini-3.5-Flash, and Qwen-3.8-Max) to quantify motion fidelity and temporal continuity. Concurrently, we record caption length (Len, in words) and shot structure match rate (SMR). "N/S" denotes outputs failing to follow the requested multi-shot structure. For these models, Motion Accuracy is evaluated without shot matching, while retaining checks on action, person, and limb correctness.

5.2 Evaluating Existing Models on FlexBench

As shown in Table 1, substantive-action scores are consistently lower than non-substantive scores across evaluated models: Qwen-3.8-Max reaches 33.50 versus 46.88, and Gemini-3.1-Pro reaches 17.80 versus 42.21. Fine-grained limb-motion captioning thus remains challenging even for models producing rich captions. Tarsier2-Recap-7B achieves a notable 22.08 GPA at just 127.9 average words, though evaluated under relaxed shot matching. The gap between GPA_{clip} and GPA shows correct content often coexists with errors in dense captions: Qwen-3.8-Max and Seed-2.1-Pro incur deductions of 8.26 and 7.82 points, respectively.

Sequence-level scores reveal a related limitation. Seed-2.1-Pro achieves the highest TSA_{w} at 46.30, while Qwen3-32B leads the evaluated large open-weight models at 27.83. These queries require consecutive actions to be correctly represented in their corresponding shot contexts, making their joint coverage more demanding. Output structure alone does not resolve this difficulty: Qwen-3.5-Omni achieves 99.34% SMR but only 13.72 on substantive actions. Conversely, the two evaluated

Table 1: Zero-shot performance on FlexBench across SOTA VLMs. Best results are in **bold**.

Model	Motion Accuracy				Motion Sequence				Output Statistics	
	Subs.	Non-subs.	GPA	GPA_{clip}	Seq_2	Seq_3	Seq_{chain}	TSA_w	Len	SMR (%)
Proprietary Frontier APIs										
Gemini-3.1-Pro	17.80	42.21	27.56	34.27	26.53	28.18	30.04	28.78	855.7	50.27
Gemini-3.5-Flash	16.70	43.61	27.46	33.98	27.03	26.88	28.77	27.86	709.4	56.22
Seed-2.1-Pro	32.24	44.38	37.10	44.92	**47.00**	44.14	**47.31**	**46.30**	782.3	68.31
Kimi-k2.6	18.47	42.13	27.93	36.04	30.35	30.26	32.94	31.62	683.4	53.57
Qwen-3.8-Max	**33.50**	**46.88**	**38.85**	**47.11**	45.24	**45.44**	46.70	46.03	1233.3	99.18
Open-Source Large Models										
Nemotron-3-Omni-30B	5.94	41.22	20.05	25.10	14.92	16.24	15.22	15.47	310.4	64.23
Gemma-4-31B	7.63	**48.37**	23.93	26.49	13.68	15.90	17.12	16.07	424.9	78.98
Qwen3-32B	**14.20**	39.91	24.48	**33.81**	**25.95**	**25.96**	**29.70**	**27.83**	674.9	96.57
Qwen-3.8-27B	13.47	44.53	**25.89**	31.66	22.76	23.04	25.28	24.10	596.7	**98.10**
Open-Weight Video Models (7B)										
VideoLLaMA2-7B	2.01	21.47	9.79	13.89	6.84	4.62	9.68	7.59	142.4	N/S
Tarsier2-Recap-7B	**8.96**	**41.75**	**22.08**	**24.98**	**9.99**	**6.84**	**12.17**	**10.13**	127.9	N/S

Table 2: Main alignment results on FlexBench. Base models in *gray italics*; best aligned in **bold**.

Model Variant	Motion Accuracy				Motion Sequence				Output Statistics	
	Subs.	Non-subs.	GPA	GPA_{clip}	Seq_2	Seq_3	Seq_{chain}	TSA_w	Len	SMR (%)
Qwen3-8B-SFT	6.92	42.31	21.08	29.88	13.30	15.54	15.81	15.23	1634	89.7
Qwen2.5-7B-Base	*3.02*	*23.17*	*11.08*	*14.26*	*8.86*	*9.88*	*10.38*	*9.93*	*1131*	*49.0*
Qwen2.5-7B-DPO	3.92	24.22	12.04	25.05	15.13	15.96	18.39	17.01	1420	94.0
Qwen2.5-7B-cDPO	3.05	25.80	12.15	25.20	15.25	14.55	17.65	16.24	1303	96.0
Qwen2.5-7B-IPO	2.94	27.14	12.62	**26.40**	**17.83**	**18.42**	19.22	**18.70**	1521	**98.7**
Qwen2.5-7B-SimPO	1.39	**30.67**	13.10	25.42	15.32	15.32	17.71	16.52	1561	97.8
Qwen2.5-7B-GM-DPO (Ours)	**6.14**	25.95	**14.06**	25.82	17.05	17.31	**19.72**	18.46	1506	98.2
Qwen3-8B-Base	*5.85*	*40.25*	*19.61*	*25.67*	*13.09*	*12.75*	*15.99*	*14.44*	*1391*	*37.2*
Qwen3-8B-DPO	7.16	38.40	19.66	28.86	17.69	19.94	20.52	19.78	1572	55.6
Qwen3-8B-cDPO	6.44	33.76	17.37	27.37	16.73	17.32	20.25	18.67	1514	63.7
Qwen3-8B-IPO	7.05	39.33	19.96	29.08	18.35	19.19	21.74	20.30	1563	86.3
Qwen3-8B-SimPO	6.64	39.58	19.82	28.60	18.84	19.77	21.35	20.37	1615	77.7
Qwen3-8B-GM-DPO (Ours)	**9.30**	**43.69**	**23.06**	**30.52**	**21.81**	**20.58**	**23.60**	**22.34**	1610	**89.8**
LLaVA-2-8B-Base	*4.24*	*48.70*	*22.02*	*23.21*	*4.94*	*5.79*	*6.38*	*5.92*	*605*	*2.2*
LLaVA-2-8B-DPO	10.66	40.73	22.69	29.15	21.17	22.53	23.33	22.66	1320	25.3
LLaVA-2-8B-cDPO	6.35	45.46	21.99	25.54	13.44	13.58	14.90	14.21	1298	36.9
LLaVA-2-8B-IPO	10.24	41.40	22.70	29.22	22.89	23.00	24.01	23.48	1413	39.5
LLaVA-2-8B-SimPO	6.93	**47.42**	23.13	26.56	16.90	18.19	15.71	16.69	1400	32.6
LLaVA-2-8B-GM-DPO (Ours)	**13.89**	42.66	**25.40**	**31.77**	**25.87**	**25.49**	**27.04**	**26.34**	1405	**44.7**

7B video models do not produce the requested shot structure. Together, these results motivate improving limb-motion fidelity within complete, shot-organized captions.

5.3 Main Alignment Results

We next examine whether targeted preference optimization can address these limitations while retaining dense caption output. As shown in Table 2, we compare preference objectives across three backbones with the Qwen3-8B SFT baseline serving as a reference for caption imitation alone.

Fine-Grained Action Grounding. Standard DPO treats minor omissions identically to fatal chirality inversions, resulting in poor performance under evaluation systems that penalize hallucinations. GM-DPO incorporates this information through severity-dependent margins and weights, achieving the highest Substantive scores among the evaluated preference methods on all three backbones. Relative to DPO, scores increase from 3.92 to 6.14 on Qwen2.5-7B, 7.16 to 9.30 on Qwen3-8B, and 10.66 to 13.89 on LLaVA-2-8B. Non-substantive scores also improve over DPO on every backbone. Although other objectives obtain higher non-substantive scores on Qwen2.5-7B and LLaVA-2-8B, GM-DPO consistently leads on the substantive actions targeted by our preference construction.

Robustness under Penalty-Aware Evaluation. To examine whether these improvements persist when incorrect assertions incur penalties, we compare the aggregate GPA scores. GM-DPO ranks first among the evaluated alignment methods on every backbone, reaching 14.06, 23.06, and 25.40, respectively, with gains of 2.02, 3.40, and 2.71 points over DPO. These gains contain two measurable contributions: GPA_{clip} increases by 0.77, 1.02, and 2.62 points, while the aggregate penalty deduction, $GPA_{clip} - GPA$, decreases from 13.01 to 11.76, 9.20 to 7.46, and 6.46 to 6.37. The improvement therefore combines higher positive credit with smaller penalties, with their relative

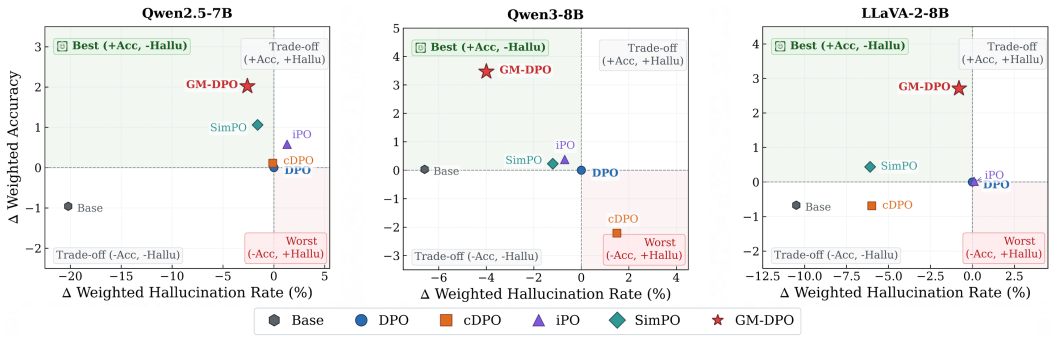


Figure 5: The accuracy-hallucination trade-off across three VLM architectures on FlexBench, normalized relative to standard DPO as origin (0, 0).

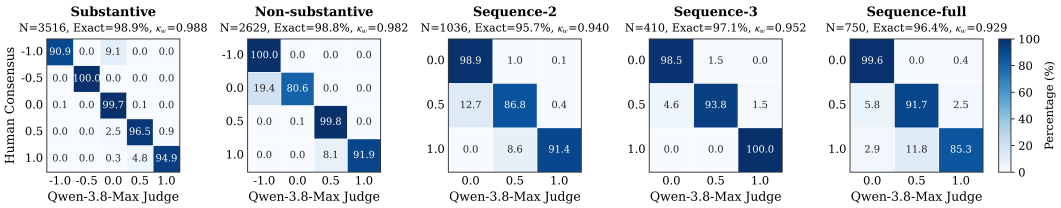


Figure 6: Row-normalized agreement matrices between human consensus and LLM.

contributions varying across backbones. In particular, the LLaVA-2-8B gain primarily reflects increased positive credit rather than a large reduction in penalties.

Consecutive-Action Coverage across Shots. Sequence queries jointly evaluate consecutive substantive actions in their corresponding shot contexts, connecting individual action fidelity with coverage across shots. Compared to DPO, GM-DPO consistently improves $\mathrm{TSA}_{\mathrm{w}}$ by 1.45, 2.56, and 3.68 points, respectively. It achieves the highest $\mathrm{Seq}_{\mathrm{chain}}$ on all three backbones and leads all sequence metrics on Qwen3-8B and LLaVA-2-8B. On Qwen2.5-7B, IPO retains a slightly higher $\mathrm{TSA}_{\mathrm{w}}$ of 18.70. These results support improved joint coverage of consecutive actions; because the queries share action requirements, they do not isolate temporal reasoning from action accuracy.

Maintaining Caption Length and Shot Structure. A common failure mode in preference optimization is length or format collapse, where models generate truncated answers to avoid penalties. As shown in Table 2, all preference-tuned variants preserve descriptive length while improving SMR over base models. Notably, on Qwen3-8B, GM-DPO matches full-SFT in both length (1610 vs. 1634) and structure compliance (89.8% vs. 89.7% SMR), demonstrating that our auxiliary NLL loss prevents linguistic collapse without standalone SFT. Nevertheless, final SMR gains remain largely bounded by each backbone's inherent instruction-following capacity.

5.4 Diagnostic Analysis and Evaluator Fidelity

Physical Fidelity and Hallucination. As shown in Figure 5, we compare changes in GPA and WHR relative to DPO to examine how physical alignment improvements relate to hallucination penalties. GM-DPO occupies the upper-left region on all three backbones, achieving the largest GPA increase among the compared alignment methods while also reducing WHR. Several baselines also reach this region, so the advantage is not exclusive quadrant membership. For example, SimPO achieves a larger WHR reduction on LLaVA-2-8B, but its GPA gain is smaller than GM-DPO's (0.44 versus 2.71 points). On Qwen3-8B, GM-DPO reduces WHR from 18.8% to 14.8%, a reduction of 4.0 percentage points, while increasing GPA by 3.40 points. These results show that improved penalty-aware fidelity can accompany reduced hallucination rates, with the balance varying across objectives and backbones. Representative caption comparisons are provided in Appendix C.4.

Evaluator Reliability and Human Agreement. To validate the reliability of FlexBench's automated evaluation protocol, we benchmark the LLM judge (`Qwen-3.8-Max`, $T = 0.0$) against consensus annotations from 20 trained experts across 8,341 query instances generated by 28 diverse models. As shown in Figure 6, exact agreement spans from 95.7% to 98.9% across all five suites, accompanied by outstanding quadratic weighted κ_w values ranging from 0.929 to 0.988. By demon-

strating robust alignment across a broad spectrum of model outputs, these results confirm the high fidelity, objectivity, and practical viability of FlexBench's automated protocol as a reliable proxy for human judgment in fine-grained video evaluation.

5.5 Ablation on GM-DPO Loss Formulation

We ablate the severity-dependent weight (ω) and margin (m), recovering DPO when both are disabled while retaining the auxiliary NLL loss. As shown in Table 3, the margin alone improves GPA by 1.16, 0.41, and 1.86 points on Qwen2.5, Qwen3, and LLaVA-2, respectively, outperforming weighting alone on every backbone. Weighting alone also improves GPA, but its benefits do not consistently extend to credited content or sequence coverage. Combining both components yields the highest GPA, $\mathrm{GPA}_{\mathrm{clip}}$, and $\mathrm{TSA}_{\mathrm{w}}$ in every ablation group, supporting their complementary roles in improving action fidelity and consecutive-action coverage. This benefit is clearest on Qwen3, where joint optimization reaches 23.06 GPA, compared with 19.93 for weighting alone and 20.07 for the margin alone. Detailed results appear in Appendix C.2.

Table 3: Ablation across backbones.

Model	ω	m	GPA	$\mathrm{GPA}_{\mathrm{clip}}$	$\mathrm{TSA}_{\mathrm{w}}$
Qwen2.5	✗	✗	12.04	25.05	17.01
	✓	✗	12.29	25.00	16.52
	✗	✓	13.20	25.44	17.73
	✓	✓	**14.06**	**25.82**	**18.46**
Qwen3	✗	✗	19.66	28.86	19.78
	✓	✗	19.93	29.42	19.88
	✗	✓	20.07	29.82	20.03
	✓	✓	**23.06**	**30.52**	**22.34**
LLaVA-2	✗	✗	22.69	29.15	22.66
	✓	✗	23.77	28.97	23.27
	✗	✓	24.55	30.63	25.27
	✓	✓	**25.40**	**31.77**	**26.34**

5.6 Generalization to Downstream Multimodal Benchmarks

To examine whether fine-grained motion alignment retains broader multimodal capabilities, we evaluate the aligned checkpoints on Video-MME (Fu et al., 2025), POPE (Li et al., 2023), and Video-ChatGPT (Maaz et al., 2024). As shown in Table 4, GM-DPO achieves the highest POPE accuracy and F1, together with the highest Video-ChatGPT overall score, among the compared alignment methods on all three backbones. It also leads Video-MME on Qwen2.5 and LLaVA-2; on Qwen3, its accuracy of 62.72 is slightly below cDPO's 62.73. Crucially, cross-benchmark comparison reveals an overarching diagnostic limitation of prevailing suites: performance deltas across disparate alignment objectives remain heavily compressed (fluctuating within ± 0.1 on Video-ChatGPT), as coarse event recognition and multiple-choice probes lack sensitivity to granular physical dynamics. In stark contrast, FlexBench exposes decisive gaps across these identical models (spanning up to 3.40 GPA points and a 21.3% WHR delta in Section 5.4), confirming metric saturation on standard benchmarks and highlighting FlexBench as an indispensable diagnostic testbed. Comprehensive breakdowns are provided in Appendix C.3.

Table 4: Cross-benchmark generalization across backbones.

Method Variant	V-MME Acc	POPE Acc	POPE F1	V-GPT Score
Q2.5-DPO	60.06	87.43	86.06	2.50
Q2.5-cDPO	59.58	87.46	86.06	2.50
Q2.5-IPO	60.45	87.44	86.09	2.46
Q2.5-SimPO	60.26	87.41	86.02	2.48
Q2.5-GM-DPO	**60.62**	**87.56**	**86.15**	**2.51**
Q3-DPO	62.35	89.37	88.98	2.70
Q3-cDPO	**62.73**	89.37	88.98	2.73
Q3-IPO	62.07	89.37	88.96	2.66
Q3-SimPO	62.24	89.35	88.94	2.64
Q3-GM-DPO	62.72	**89.40**	**89.02**	**2.75**
L2-DPO	64.67	88.78	88.31	2.51
L2-cDPO	65.14	88.88	88.34	2.56
L2-IPO	65.19	88.77	88.11	2.56
L2-SimPO	64.78	88.49	88.04	2.55
L2-GM-DPO	**65.39**	**88.97**	**88.41**	**2.57**

6 Conclusion

We presented a framework combining fine-grained evaluation and severity-aware preference learning to improve limb-motion fidelity in dense video captions. FlexBench systematically evaluates each identifiable person's actions and limb states across shots, while GPA distinguishes credited content from incorrect claims through graded scores and penalties. GM-DPO brings this rubric into preference margins and loss weights, consistently improving substantive-action fidelity and consecutive-action coverage over DPO across three backbones and training seeds. These gains accompany sustained long-form output and competitive performance on broader multimodal benchmarks, supporting targeted motion alignment within general-purpose captioning models. However, monocular ambiguity remains a challenge, and these results do not establish the accuracy of other caption content, such as camera angles, camera motion, or sound. Future work will evaluate these dimensions and examine whether more accurate motion captions provide better supervision for downstream learning. Further limitations and research directions are discussed in Appendix D.

Reproducibility Statement

We provide a supplementary ZIP archive containing data-construction scripts, GM-DPO training code, evaluation code, prompts, and data subsets for running the experimental pipeline. The archive also includes the FlexBench annotations and evaluation checklists, together with the corresponding evaluation implementation. A detailed README provides step-by-step instructions for preparing the data, running training, and evaluating model outputs. The main paper and appendices document the grading rubric, preference-pair construction, baseline objectives, training configurations, and random-seed settings, with implementation details provided in Appendix B. Due to supplementary file size limits, the full video data and trained model weights will be released following publication, subject to the applicable source licenses and usage conditions.

AI Use Statement

We used generative AI tools for video-caption pre-annotation, synthetic perturbation generation for preference training, and automated caption evaluation, as described in the paper. VLM-generated benchmark annotations underwent human refinement and review, and automated evaluation was assessed against human judgments. We also used AI assistance for literature search and synthesis, and manuscript organization and language editing. The authors are responsible for verifying the AI-assisted material and for the final text, methods, claims, and research artifacts.

Ethics Statement

This work studies video captioning using existing video sources. Human contributors participated in annotation, quality review, and caption evaluation, including the blinded perturbation study described in the appendix. Person identifiers are used to associate visible actions across shots. Any release of source videos will respect applicable licenses and usage conditions.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4447–4455, 2024.

ByteDance Seed Team. Seed-2.1-Pro: Multimodal foundation api for continuous visual understanding. `https://www.volcengine.com/product/seed`, 2026.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026. URL `https://arxiv.org/abs/2601.10611`.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Jiashuo Sun, Shuhuai Ren, Renrui Zhang, Ning Wang, Bin Wang, Ruoyu Wang, Yan Chen, Runpeng Yu, Liang Chen, Dong Shen, Shaohao Lu, Ming-Hsuan Yang, Caifeng Shan, and Xinglong Shen. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Google Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/`, 2026a.

Google Gemini Team. Gemini 3.5: Frontier intelligence with action. `https://blog.google/technology/ai/google-gemini-3-5/`, 2026b.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6047–6056, 2018.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dahun Kim, AJ Piergiovanni, Ganesh Satish Mallya, and Anelia Angelova. VideoComp: Advancing fine-grained compositional and temporal alignment in video-text models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kimi Team and Moonshot AI. Kimi K2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Ji Soo Lee, Jongha Kim, Jeehye Na, Jinyoung Park, and Hyunwoo J. Kim. VidChain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning. *arXiv preprint arXiv:2501.06761*, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023.

Boda Lin, Yongjie Zhu, Xiaocheng Gong, Wenyu Qin, and Meng Wang. KPM-Bench: A kinematic parsing motion benchmark for fine-grained motion-centric video understanding. *arXiv preprint arXiv:2602.17768*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand video time? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13745–13768, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12586–12602, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 40645–40661, 2023.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 124430–124458, 2024.

Eric Mitchell. A note on DPO with noisy preferences & relationship to IPO. Technical note, 2023. URL https://ericmitchell.ai/cdpo.pdf.

Qwen Team. Qwen3.5-Omni technical report, 2026a. URL https://arxiv.org/abs/2604.15804.

Qwen Team. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability, 2026b. URL https://arxiv.org/abs/2608.30320.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. *arXiv preprint arXiv:2503.14935*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, 2024.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

Appendix Overview

The appendix is organized into four sections. Section A details FlexBench video selection, annotation, and quality assurance, together with perturbation protocols and human assessment of caption naturalness. Section B documents the training setup, random-seed settings, baseline objectives, and method-specific hyperparameters. Section C begins with results across three random seeds (Section C.1), followed by complete ablation results (Section C.2), evaluations on Video-MME, POPE, and Video-ChatGPT (Section C.3), and qualitative caption comparisons (Section C.4). Section D discusses limitations and future directions, including extensions of the grading rubric, ambiguity in monocular video, and broader evaluation of dense caption content.

A FlexBench Curation and Annotation Details

A.1 Video Filtering and Sequence Curation

Raw footage harvested from the AVA repository (Gu et al., 2018) and YouTube Creative Commons streams was re-segmented and concatenated to build an initial pool of over 56k candidate sequences. We applied automated heuristic screening to filter out sequences containing shots shorter than 1.5 s, instances with minimal physical motion, scenery-dominated sequences, and sensitive or unsafe footage, reducing the pool to 12k sequences. Subsequently, domain experts conducted stratified manual selection prioritizing scenario diversity, actor density, and rich limb-level physical interactions. This curation yielded the final FlexBench benchmark comprising 660 verified video sequences spanning 3,105 distinct continuous shots.

A.2 Human-in-the-Loop Annotation and Multi-Round Verification

Following automated VLM pre-captioning, an expert panel of 18 trained annotators holding bachelor's degrees refined candidate annotations under a standardized rubric covering cross-shot subject tracking, visual appearance, interacted affordances, limb chirality (explicitly tagging left, right, or bimanual execution), and chronological connectives. Annotations underwent three rounds of blind cross-verification. Raw inter-annotator agreement was 94.2%, with Fleiss' $\kappa = 0.89$. For the remaining edge cases—primarily involving extreme perspective foreshortening or partial self-occlusions—chirality and action states were finalized via panel consensus arbitration based on a two-thirds majority voting protocol under multi-frame zoom examination. Finally, atomic positive QA probes were derived from the verified captions, with every probe independently audited across two rounds of human inspection to ensure precise physical factuality.

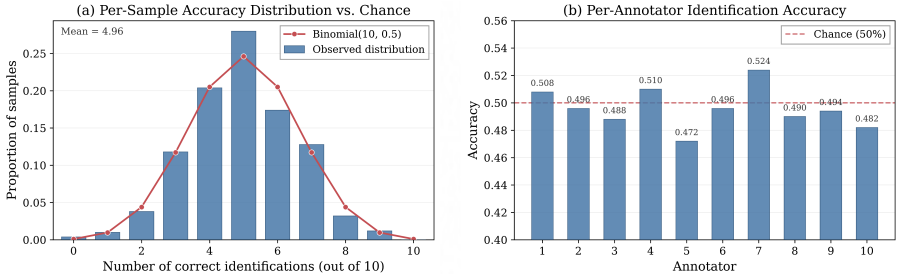


Figure 7: Double-blind human identification study on preference pair naturalness. (a) The empirical distribution of correct identifications per sample closely matches the theoretical Binomial$(10, 0.5)$ chance baseline. (b) Individual identification accuracies across all 10 human evaluators fluctuate tightly around the 50% chance threshold (overall accuracy = 49.60%, binomial $p = 0.5813 > 0.05$).

A.3 Linguistic Naturalness and Non-Spuriousness of Preference Pairs

Locally Constrained Perturbation Protocols. To eliminate textual distribution discrepancies between preferred and dispreferred pairs, all perturbations are strictly confined to local limb kinematics, keeping the surrounding sentence structures, character appearances, and stylistic register completely frozen. As introduced in Section 4.4, directional laterality is modified via deterministic token substitution (e.g., swapping *"left"* and *"right"* hands or stripping laterality qualifiers), while temporal omissions, subtle subaction variations, and phantom actions are synthesized using Gemini-3.5-Flash under rigid schema constraints. Because ground-truth captions (y_w) are originally initialized by frontier VLMs before human refinement, both y_w and y_l share identical underlying linguistic foundations and syntactic distributions.

Double-Blind Human Indistinguishability Evaluation. To verify that models optimize genuine physical grounding rather than discriminating between human-edited and LLM-synthesized writing styles, we conducted a double-blind identification experiment. Ten independent annotators with no prior exposure to the dataset evaluated 500 randomly sampled blind pairs (5,000 total judgments) to identify human-refined versus LLM-modified descriptions. As shown in Figure 7, human evaluators achieved an overall accuracy of 49.60% (virtually identical to random chance; binomial test $p = 0.5813 > 0.05$), with both sample-level and individual identification distributions conforming closely to the theoretical Binomial$(10, 0.5)$ null distribution. This confirms that our targeted kinematic modifications introduce no stylistic artifacts or distributional drift, ensuring that GM-DPO is driven strictly by visual-physical motion perception.

B Experimental Setup and Baseline Objectives

All preference alignment experiments are conducted using the `ms-swift` framework. We perform full-parameter instruction tuning on the language backbones while freezing both the vision encoders and cross-modal projector modules. Training is executed in `bfloat16` precision using the AdamW optimizer with a learning rate of 5×10^{-7}. The global batch size is 128. Across all alignment objectives, an auxiliary supervised next-token prediction loss ($+1.0 \times \mathcal{L}_{\text{NLL}}$) is incorporated to preserve generation fluency. Given preference tuples $\mathcal{D} = (x, y_w, y_l)$, Table 5 summarizes the mathematical formulations and key hyperparameter configurations across the evaluated offline alignment methods. For both SFT and preference optimization, we conduct three training runs with $(\texttt{seed}, \texttt{data_seed})$ set to $(42, 42)$, $(106, 106)$, and $(298, 298)$. The global seed controls stochastic training operations, while `data_seed` controls data shuffling. Both seeds vary together across runs, and reported results for trained models are arithmetic means over the three runs.

Table 5: Mathematical formulations and key hyperparameter configurations of benchmarked offline preference optimization objectives. Let $\Delta r_\theta = \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$ denote the implicit reward margin. All methods incorporate an auxiliary $+1.0 \times \mathcal{L}_{\text{NLL}}$ loss.

Method	Preference Loss ℓ_M	Hyperparameters
DPO (Rafailov et al., 2023)	$-\log \sigma(\Delta r_\theta)$	$\beta = 0.1$
cDPO (Mitchell, 2023)	$-(1-\varepsilon)\log \sigma(\Delta r_\theta) - \varepsilon \log \sigma(-\Delta r_\theta)$	$\beta = 0.1,\ \varepsilon = 0.1$
IPO (Azar et al., 2024)	$\left(\bar{h}_\theta - \frac{1}{2\beta}\right)^2$	$\beta = 0.5$
SimPO (Meng et al., 2024)	$-\log \sigma(\beta[q_\theta(y_w \mid x) - q_\theta(y_l \mid x)] - \gamma_{\text{S}})$	$\beta = 2.0,\ \gamma_{\text{S}} = 1.0$
GM-DPO (Ours)	$-(1+\alpha\Delta S)\log \sigma(\Delta r_\theta - \gamma \Delta S)$	$\beta = 0.1,\ \gamma = 0.2,\ \alpha = 0.1$

Table 6: FlexBench results across three training seeds, reported as mean ± sample standard deviation. The highest mean within each backbone is highlighted in **bold**.

Method	Motion Accuracy				Motion Sequence			
	Subs.	Non-subs.	GPA	GPA_{clip}	Seq_2	Seq_3	$\text{Seq}_{\text{chain}}$	TSA_{w}
Qwen2.5-7B								
DPO	3.92 ± 0.12	24.22 ± 0.29	12.04 ± 0.11	25.05 ± 0.29	15.13 ± 0.34	15.96 ± 0.13	18.39 ± 0.21	17.01 ± 0.18
cDPO	3.05 ± 0.33	25.80 ± 0.13	12.15 ± 0.25	25.20 ± 0.29	15.25 ± 0.45	14.55 ± 0.33	17.65 ± 0.25	16.24 ± 0.13
IPO	2.94 ± 0.33	27.14 ± 0.14	12.62 ± 0.18	$\mathbf{26.40} \pm 0.27$	$\mathbf{17.83} \pm 0.39$	$\mathbf{18.42} \pm 0.33$	19.22 ± 0.20	$\mathbf{18.70} \pm 0.25$
SimPO	1.39 ± 0.39	$\mathbf{30.67} \pm 0.13$	13.10 ± 0.20	25.42 ± 0.18	15.32 ± 0.25	15.32 ± 0.22	17.71 ± 0.42	16.52 ± 0.25
GM-DPO	$\mathbf{6.14} \pm 0.27$	25.95 ± 0.31	$\mathbf{14.06} \pm 0.05$	25.82 ± 0.22	17.05 ± 0.31	17.31 ± 0.17	$\mathbf{19.72} \pm 0.31$	18.46 ± 0.25
Qwen3-8B								
SFT	6.92 ± 0.29	42.31 ± 0.14	21.08 ± 0.14	29.88 ± 0.12	13.30 ± 0.14	15.54 ± 0.20	15.81 ± 0.35	15.23 ± 0.23
DPO	7.16 ± 0.33	38.40 ± 0.46	19.66 ± 0.31	28.86 ± 0.24	17.69 ± 0.36	19.94 ± 0.28	20.52 ± 0.24	19.78 ± 0.14
cDPO	6.44 ± 0.36	33.76 ± 0.14	17.37 ± 0.16	27.37 ± 0.39	16.73 ± 0.17	17.32 ± 0.14	20.25 ± 0.23	18.67 ± 0.08
IPO	7.05 ± 0.15	39.33 ± 0.42	19.96 ± 0.24	29.08 ± 0.18	18.35 ± 0.25	19.19 ± 0.34	21.74 ± 0.25	20.30 ± 0.12
SimPO	6.64 ± 0.29	39.58 ± 0.41	19.82 ± 0.30	28.60 ± 0.18	18.84 ± 0.29	19.77 ± 0.18	21.35 ± 0.21	20.37 ± 0.21
GM-DPO	$\mathbf{9.30} \pm 0.30$	$\mathbf{43.69} \pm 0.28$	$\mathbf{23.06} \pm 0.16$	$\mathbf{30.52} \pm 0.21$	$\mathbf{21.81} \pm 0.46$	$\mathbf{20.58} \pm 0.38$	$\mathbf{23.60} \pm 0.37$	$\mathbf{22.34} \pm 0.17$
LLaVA-2-8B								
DPO	10.66 ± 0.31	40.73 ± 0.20	22.69 ± 0.19	29.15 ± 0.28	21.17 ± 0.26	22.53 ± 0.24	23.33 ± 0.35	22.66 ± 0.23
cDPO	6.35 ± 0.36	45.46 ± 0.39	21.99 ± 0.34	25.54 ± 0.15	13.44 ± 0.29	13.58 ± 0.31	14.90 ± 0.40	14.21 ± 0.30
IPO	10.24 ± 0.34	41.40 ± 0.28	22.70 ± 0.17	29.22 ± 0.28	22.89 ± 0.40	23.00 ± 0.38	24.01 ± 0.28	23.48 ± 0.23
SimPO	6.93 ± 0.27	$\mathbf{47.42} \pm 0.09$	23.13 ± 0.15	26.56 ± 0.16	16.90 ± 0.17	18.19 ± 0.33	15.71 ± 0.32	16.69 ± 0.23
GM-DPO	$\mathbf{13.89} \pm 0.24$	42.66 ± 0.32	$\mathbf{25.40} \pm 0.15$	$\mathbf{31.77} \pm 0.07$	$\mathbf{25.87} \pm 0.29$	$\mathbf{25.49} \pm 0.44$	$\mathbf{27.04} \pm 0.27$	$\mathbf{26.34} \pm 0.22$

C Comprehensive Experimental Results and Visualizations

C.1 Robustness across Random Seeds

We repeat each preference method on all three backbones and SFT on Qwen3-8B using the three seed configurations in Appendix B, jointly varying the training and data-order seeds while fixing all other settings. Table 6 reports means and sample standard deviations ($n = 3$, denominator $n - 1$), with the highest mean per backbone in bold; the means match Table 2. GM-DPO leads the evaluated preference methods in substantive-action accuracy, GPA, and $\text{Seq}_{\text{chain}}$ on every backbone in all three runs. It also consistently leads in GPA_{clip} and TSA_{w} on Qwen3-8B and LLaVA-2-8B, and exceeds SFT on Qwen3-8B in mean substantive-action accuracy (9.30 versus 6.92) and GPA (23.06 versus 21.08). On Qwen2.5-7B, IPO retains higher mean GPA_{clip} and TSA_{w}, with the TSA ranking between IPO and GM-DPO varying across runs.

Paired Improvements over DPO. We pair runs with identical training and data-order seed settings and compute the within-pair GPA difference d_i between GM-DPO and DPO. Table 7 reports these gains and pointwise 95% confidence intervals (CIs), calculated as $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$, where s_d is the sample standard deviation of the paired differences. The intervals use two degrees of freedom and no multiple-comparison adjustment. All paired gains are pos-

Table 7: Paired GPA gains over DPO.

Backbone	GPA Gain			Mean	95% CI
	Run 1	Run 2	Run 3		
Qwen2.5-7B	+2.12	+1.83	+2.11	+2.02	[1.61, 2.43]
Qwen3-8B	+3.41	+3.80	+2.99	+3.40	[2.39, 4.41]
LLaVA-2-8B	+2.57	+2.67	+2.90	+2.71	[2.29, 3.13]

Table 8: Comprehensive ablation study of GM-DPO core components across all backbones on FlexBench. Omitting both components recovers the standard DPO baseline. Best performance within each model family is in **bold**.

Model	Weight	Margin	Motion Accuracy (Penalty-Aware)				Motion Sequence (Non-Penalty)			
			Subs.	Non-subs.	GPA	$\mathrm{GPA}_{\mathrm{clip}}$	Seq_2	Seq_3	$\mathrm{Seq}_{\mathrm{chain}}$	$\mathrm{TSA}_{\mathrm{w}}$
Qwen2.5-7B	✗	✗	3.92	24.22	12.04	25.05	15.13	15.96	18.39	17.01
	✓	✗	4.19	24.44	12.29	25.00	15.86	14.61	17.92	16.52
	✗	✓	5.66	24.51	13.20	25.44	16.16	15.89	19.47	17.73
	✓	✓	**6.14**	**25.95**	**14.06**	**25.82**	**17.05**	**17.31**	**19.72**	**18.46**
Qwen3-8B	✗	✗	7.16	38.40	19.66	28.86	17.69	19.94	20.52	19.78
	✓	✗	8.19	37.53	19.93	29.42	19.25	18.98	20.67	19.88
	✗	✓	8.76	37.04	20.07	29.82	19.40	19.18	20.79	20.03
	✓	✓	**9.30**	**43.69**	**23.06**	**30.52**	**21.81**	**20.58**	**23.60**	**22.34**
LLaVA-2-8B	✗	✗	10.66	40.73	22.69	29.15	21.17	22.53	23.33	22.66
	✓	✗	10.23	**44.09**	23.77	28.97	21.52	23.66	23.73	23.27
	✗	✓	11.88	43.55	24.55	30.63	24.81	25.00	25.62	25.27
	✓	✓	**13.89**	42.66	**25.40**	**31.77**	**25.87**	**25.49**	**27.04**	**26.34**

itive, and all intervals lie above zero. These intervals summarize cross-run uncertainty on the fixed evaluation set under independent, approximately normal paired differences. With only three runs, this assumption cannot be reliably checked; the intervals therefore provide supplementary evidence alongside the consistently positive observed gains.

C.2 Full Fine-Grained Kinematic and Sequential Ablation

Table 8 extends the main-paper ablation to all action and sequence metrics. We separately evaluate the severity-dependent margin and loss weight, retaining the same auxiliary NLL term across variants. Disabling both components recovers the DPO baseline.

Effect of the Severity-Dependent Margin. The margin-only variant improves substantive-action scores from 3.92 to 5.66 on Qwen2.5-7B, 7.16 to 8.76 on Qwen3-8B, and 10.66 to 11.88 on LLaVA-2-8B. Its GPA gains over DPO are 1.16, 0.41, and 1.86 points, respectively. Improvements also extend to $\mathrm{GPA}_{\mathrm{clip}}$ and $\mathrm{TSA}_{\mathrm{w}}$ on all three backbones, indicating gains in both credited action content and consecutive-action coverage.

Effect of Severity-Dependent Weighting. Weighting alone increases GPA by 0.25, 0.27, and 1.08 points on the three backbones, but its effects on other metrics are mixed. On Qwen2.5-7B, $\mathrm{TSA}_{\mathrm{w}}$ decreases from 17.01 to 16.52; on LLaVA-2-8B, $\mathrm{GPA}_{\mathrm{clip}}$ declines from 29.15 to 28.97 despite the higher GPA. Thus, weighting alone improves penalty-aware performance without consistently increasing credited content or sequence coverage.

Combining Both Components. The full objective achieves the highest GPA, $\mathrm{GPA}_{\mathrm{clip}}$, and $\mathrm{TSA}_{\mathrm{w}}$ within every ablation group. On Qwen3-8B, joint optimization reaches 23.06 GPA, compared with 19.93 for weighting alone and 20.07 for the margin alone. On LLaVA-2-8B, it further improves $\mathrm{TSA}_{\mathrm{w}}$ from the margin-only result of 25.27 to 26.34. These results support combining both components, though joint gains vary across backbones. The ablation establishes the full formulation's benefit; isolating severity assignments from overall weighting and margin strength requires additional controls.

C.3 Extended Downstream Generalization Benchmarks

To evaluate whether preference optimization on fine-grained physical dynamics induces catastrophic forgetting or impairs general multimodal reasoning, we benchmark aligned policies across three established downstream suites: Video-MME (Fu et al., 2025) for duration-stratified video understanding, POPE (Li et al., 2023) for binary object hallucination diagnosis, and Video-ChatGPT (Maaz et al., 2024) for open-ended conversational generation.

Comprehensive Temporal Reasoning on Video-MME. As shown in Table 9, GM-DPO achieves overall accuracies of 60.62%, 62.72%, and 65.39%, improving over the corresponding base models by 1.25, 1.02, and 0.69 percentage points. It also exceeds DPO on all three duration splits for

Table 9: Generalization performance on Video-MME across video duration splits and overall scores. Reference base models are shown in *gray italics*; within each backbone family, the best result among aligned models is in **bold**, and the second-best is underlined. The overall column is shaded in blue.

Model Variant	Short	Medium	Long	Overall
Qwen2.5-7B-Base	*68.89*	*59.11*	*50.11*	*59.37*
Qwen2.5-7B-DPO	69.11	**60.18**	50.89	60.06
Qwen2.5-7B-cDPO	68.89	59.53	50.33	59.58
Qwen2.5-7B-IPO	69.56	**60.18**	51.61	60.45
Qwen2.5-7B-SimPO	69.40	60.05	51.34	60.26
Qwen2.5-7B-GM-DPO (Ours)	**69.75**	60.17	**51.93**	**60.62**
Qwen3-8B-Base	*73.00*	*59.21*	*52.88*	*61.70*
Qwen3-8B-DPO	73.33	59.60	54.11	62.35
Qwen3-8B-cDPO	73.89	**59.62**	**54.67**	**62.73**
Qwen3-8B-IPO	74.00	58.78	53.44	62.07
Qwen3-8B-SimPO	73.67	59.06	54.00	62.24
Qwen3-8B-GM-DPO (Ours)	**74.22**	59.61	54.34	62.72
LLaVA-2-8B-Base	*77.56*	*61.44*	*55.11*	*64.70*
LLaVA-2-8B-DPO	77.56	61.22	55.22	64.67
LLaVA-2-8B-cDPO	78.50	61.30	55.62	65.14
LLaVA-2-8B-IPO	78.56	61.33	55.67	65.19
LLaVA-2-8B-SimPO	78.22	60.56	55.56	64.78
LLaVA-2-8B-GM-DPO (Ours)	**78.67**	**61.68**	**55.82**	**65.39**

Table 10: Zero-shot object hallucination evaluation on POPE across adversarial, popular, and random splits. Best aligned results within each model group are highlighted in **bold**, and second-best are underlined; **Overall Acc** is shaded in light blue, and **F1** is shaded in light pink.

Model Variant	Adversarial	Popular	Random	Overall Acc	Yes%	F1
Qwen2.5-7B-Base	*86.50*	*87.37*	*88.23*	*87.37*	*40.03*	*85.97*
Qwen2.5-7B-DPO	86.53	87.43	88.33	87.43	**40.14**	86.06
Qwen2.5-7B-cDPO	86.57	87.47	88.33	87.46	40.12	86.06
Qwen2.5-7B-IPO	86.53	87.40	88.40	87.44	40.29	86.09
Qwen2.5-7B-SimPO	86.50	87.38	88.36	87.41	40.10	86.02
Qwen2.5-7B-GM-DPO (Ours)	**86.67**	**87.53**	**88.47**	**87.56**	40.12	**86.15**
Qwen3-8B-Base	*87.17*	*89.17*	*91.57*	*89.30*	*46.08*	*88.86*
Qwen3-8B-DPO	87.00	**89.30**	91.80	89.37	46.50	88.98
Qwen3-8B-cDPO	87.02	89.28	91.81	89.37	46.51	88.98
Qwen3-8B-IPO	87.01	89.28	91.81	89.37	46.48	88.96
Qwen3-8B-SimPO	87.00	89.26	91.78	89.35	46.47	88.94
Qwen3-8B-GM-DPO (Ours)	**87.07**	**89.30**	**91.83**	**89.40**	**46.53**	**89.02**
LLaVA-2-8B-Base	*86.51*	*88.23*	*90.77*	*88.50*	*44.27*	*88.10*
LLaVA-2-8B-DPO	86.72	88.50	91.13	88.78	**45.20**	88.31
LLaVA-2-8B-cDPO	**86.88**	88.61	91.16	88.88	45.11	88.34
LLaVA-2-8B-IPO	86.67	88.37	91.28	88.77	44.98	88.11
LLaVA-2-8B-SimPO	86.44	88.27	90.77	88.49	44.48	88.04
LLaVA-2-8B-GM-DPO (Ours)	86.80	**88.65**	**91.47**	**88.97**	45.16	**88.41**

every backbone. The gains vary by duration: on LLaVA-2-8B, the improvement is larger on Short videos (1.11 percentage points) than on Long videos (0.60 percentage points). These results support retained performance across video durations following motion-focused alignment.

Cross-Domain Object Hallucination Suppression on POPE. On the POPE benchmark (Table 10), GM-DPO demonstrates robust zero-shot transfer in suppressing existential object hallucinations across Adversarial, Popular, and Random subsets. GM-DPO attains the highest Overall Acc across all backbones (87.56 on Qwen2.5-7B, 89.40 on Qwen3-8B, and 88.97 on LLaVA-2-8B) alongside top balanced F1 scores (86.15, 89.02, and 88.41, respectively). Concurrently, baseline objectives such as IPO and SimPO exhibit mild degradation on LLaVA-2-8B (F1 falling to 88.11 and 88.04), suggesting that uncalibrated quadratic margins or rigid sequence-length penalties can skew response calibration under binary probes. In contrast, GM-DPO stabilizes Yes-ratio distributions around 40.12–46.53%, preventing the binary over-affirmation collapse common to multimodal alignment.

Open-Ended Video Question Answering on Video-ChatGPT. Table 11 details per-dimension evaluations on Video-ChatGPT across Generic Understanding, Temporal Reasoning, and Consis-

Table 11: Video-ChatGPT evaluation across five dimensions using Qwen-3.8-Max as the judge on a 1–5 scale. Overall is the unweighted mean of the five scores, rounded to two decimal places. Base models are shown in *gray italics*. Best and second-best aligned results within each backbone are **bold** and underlined, respectively; ties at the displayed precision share the same formatting.

Model Variant	Generic Understanding			Temporal Reasoning	Consistency	Overall
	Correctness	Detail	Context			
Qwen2.5-7B-Base	*2.33*	*2.23*	*2.63*	*2.16*	*2.91*	*2.45*
Qwen2.5-7B-DPO	**2.32**	2.33	**2.64**	2.22	2.98	2.50
Qwen2.5-7B-cDPO	2.31	**2.35**	2.63	2.23	2.98	2.50
Qwen2.5-7B-IPO	2.30	2.29	2.58	2.19	2.92	2.46
Qwen2.5-7B-SimPO	2.31	2.32	2.60	2.23	2.96	2.48
Qwen2.5-7B-GM-DPO (Ours)	**2.32**	**2.35**	2.63	**2.25**	**2.99**	**2.51**
Qwen3-8B-Base	*2.55*	*2.57*	*2.89*	*2.52*	*3.14*	*2.73*
Qwen3-8B-DPO	2.53	2.53	2.85	2.51	3.07	2.70
Qwen3-8B-cDPO	**2.57**	2.57	2.90	2.53	3.10	2.73
Qwen3-8B-IPO	2.47	2.49	2.79	2.48	3.07	2.66
Qwen3-8B-SimPO	2.45	2.48	2.79	2.45	3.04	2.64
Qwen3-8B-GM-DPO (Ours)	2.56	**2.58**	**2.91**	**2.55**	**3.15**	**2.75**
LLaVA-2-8B-Base	*2.27*	*2.35*	*2.56*	*2.36*	*2.81*	*2.47*
LLaVA-2-8B-DPO	2.31	2.45	2.56	2.41	2.84	2.51
LLaVA-2-8B-cDPO	2.33	2.54	**2.65**	2.40	2.87	2.56
LLaVA-2-8B-IPO	2.33	2.53	2.64	2.40	2.88	2.56
LLaVA-2-8B-SimPO	2.35	2.51	2.63	**2.42**	2.85	2.55
LLaVA-2-8B-GM-DPO (Ours)	**2.36**	**2.55**	2.64	**2.42**	**2.89**	**2.57**

tency suites. GM-DPO achieves the highest Overall score among the aligned variants on all three backbones: 2.51 on Qwen2.5-7B, 2.75 on Qwen3-8B, and 2.57 on LLaVA-2-8B. These correspond to improvements of 0.01, 0.05, and 0.06 over DPO, respectively. GM-DPO also leads Consistency on all three backbones, with scores of 2.99, 3.15, and 2.89, and achieves the best or tied-best Temporal Reasoning scores. The advantages are not uniform across every dimension: DPO scores higher on Qwen2.5-7B's Context dimension (2.64 versus 2.63), while cDPO scores higher on Qwen3-8B's Correctness dimension (2.57 versus 2.56). These results support competitive overall response quality, with consistent advantages in temporal reasoning and consistency among the compared alignment methods.

C.4 Qualitative Visualizations and Case Studies

Figure 8 illustrates both the relative improvement and remaining limitations of GM-DPO. It obtains the highest cumulative score in this example (+1.5), compared with -2.5 for the base model and -1.5 for DPO. However, its caption still contains two laterality errors, each assigned -0.5. The example therefore illustrates improved aggregate action fidelity rather than error-free limb grounding.

D Limitations and Future Work

Our results support incorporating action-error severity into preference learning to improve limb-motion fidelity. Despite consistent gains across three VLM backbones, several limitations warrant further investigation.

Extending the Grading Rubric to Broader Motion Domains. Our five-level action-grading rubric distinguishes correct and partially specified content, omissions, limb errors, and fabricated actions, yielding four perturbation-severity levels for training. It targets identifiable limb actions, including bimanual coordination and laterality, with explicit human-annotated reference facts. Extending this rubric to continuous, multi-phase activities requires additional criteria for whole-body transitions, movement speed, and deformable object interactions. Such activities may involve overlapping errors that a single perturbation category cannot capture. Future work could investigate richer grading schemes and their agreement with human judgments while preserving the distinction between incomplete coverage and incorrect assertions.

Ambiguity in Monocular Video. FlexBench and our training pipeline use monocular videos spanning diverse scenes and viewpoints. These inputs lack explicit depth measurements, while self-

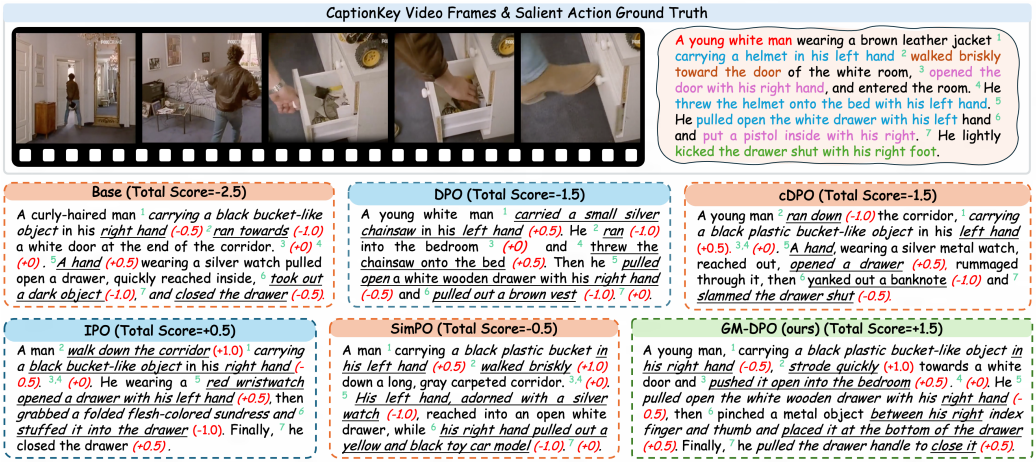


Figure 8: Qualitative comparison of physical motion understanding on a representative multi-shot sequence from FlexBench. Predictions highlight salient limb manipulation and chronological interactions across 7 atomic checkpoints, with local premise scores annotated in parentheses (-1.0 for confabulations, -0.5 for chirality/imprecise errors, $+0.5/+1.0$ for grounded dynamics). Each block header reports the unnormalized raw cumulative score (Total Score $= \sum_{i=1}^{7} s_i$).

occlusion, extreme viewing angles, and camera cuts can obscure limb identity and motion. Such ambiguities may contribute to residual errors, although our experiments do not isolate their effects from limitations in model perception or caption generation. Skeletal representations or multi-view geometric priors may help resolve ambiguous interactions. Evaluating these approaches would require distinguishing recoverable information from details that remain unobservable.

Scope of Dense Caption Evaluation. Our evaluation focuses on limb-motion fidelity, per-person coverage, and shot structure. Caption length and structural compliance characterize the output but do not establish the accuracy or completeness of other content. Although experiments on three additional benchmarks provide evidence of broader multimodal performance, they do not directly assess camera angles, camera motion, scene details, or sound descriptions within dense captions. Dedicated evaluation of these dimensions, including audio-grounded assessment when audio inputs are available, remains future work.