Omni-Reward: Towards Generalist Omni-Modal Reward Modeling with Free-Form Preferences

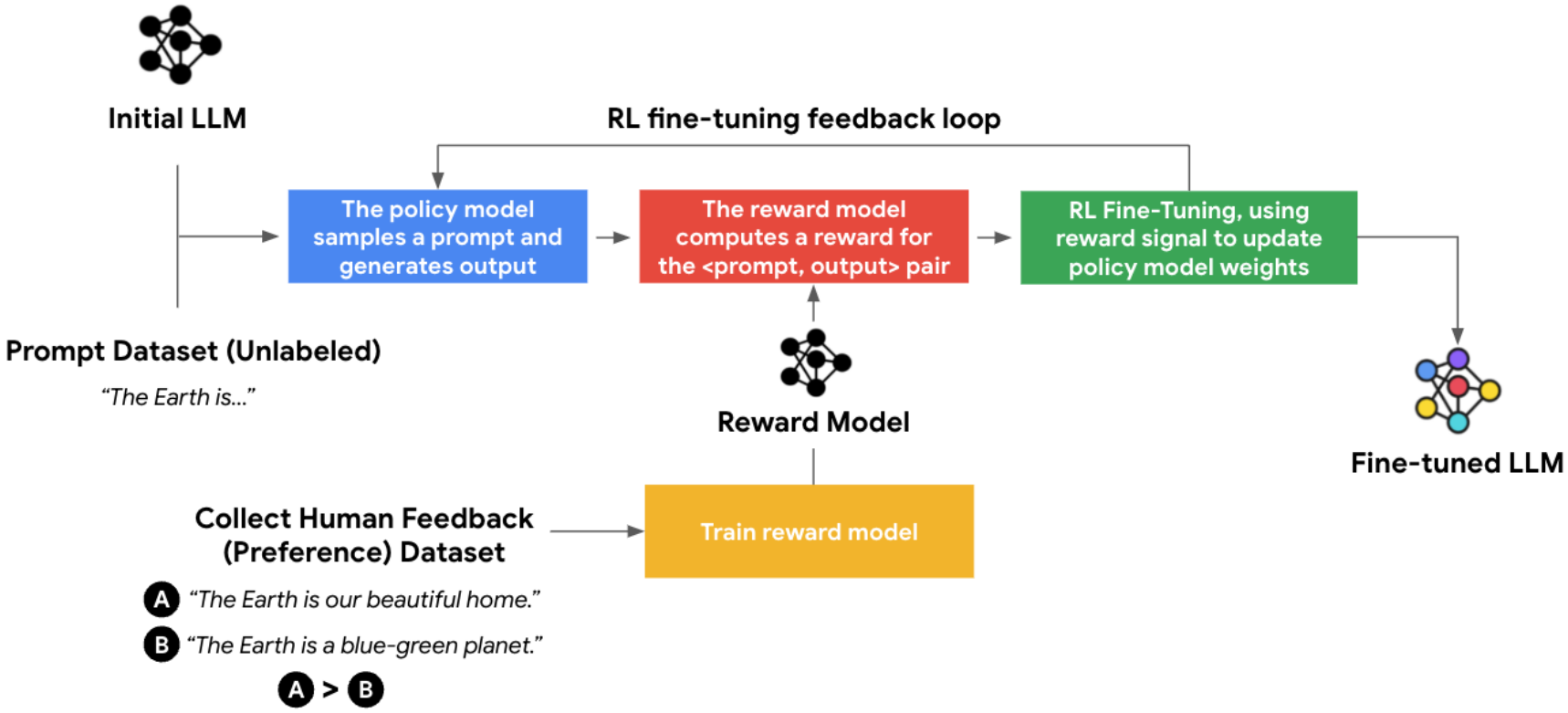
Zhuoran Jin*, Hongbang Yuan*, Kejian Zhu*,
Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao

The Fourteenth International Conference on Learning Representations
April, 2026

Benchmark: https://hf.co/datasets/HongbangYuan/OmniRewardBench
Dataset: https://hf.co/datasets/jinzhuoran/OmniRewardData
Model: https://hf.co/jinzhuoran/OmniRewardModel
Code: https://github.com/HongbangYuan/OmniReward

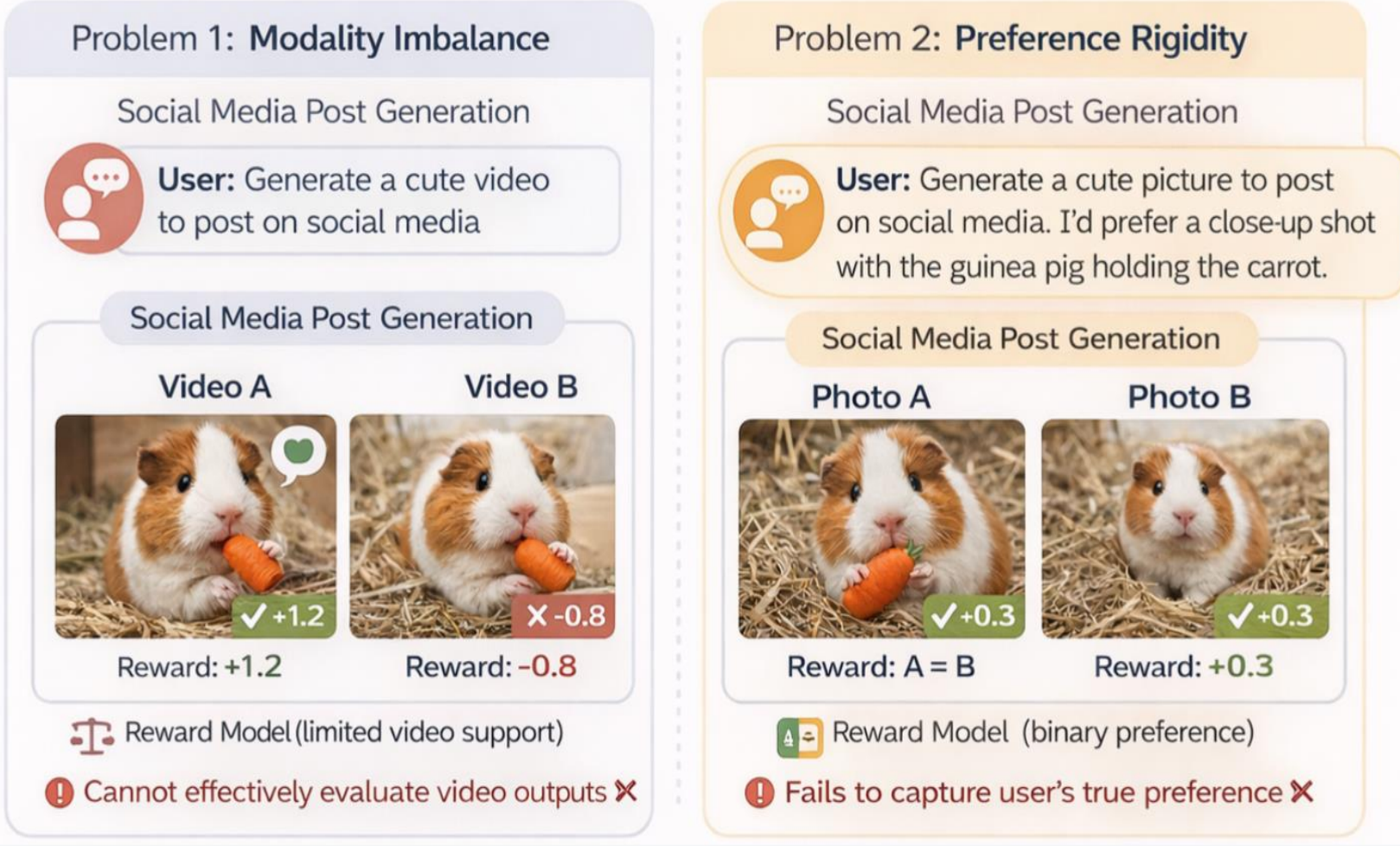
What is a Reward Model?

- A reward model serves as a **learned proxy** for human preferences, assigning scores to model outputs to reflect how well they align with human expectations, and providing feedback signals to guide model optimization in **RLHF**



Challenges of Reward Modeling

- **Modality Imbalance**: limited support beyond text and images, limiting performance on video, audio, and 3D tasks
- **Preference Rigidity**: reliance on fixed binary preferences that fail to capture diverse and fine-grained human judgments



Omni-Reward

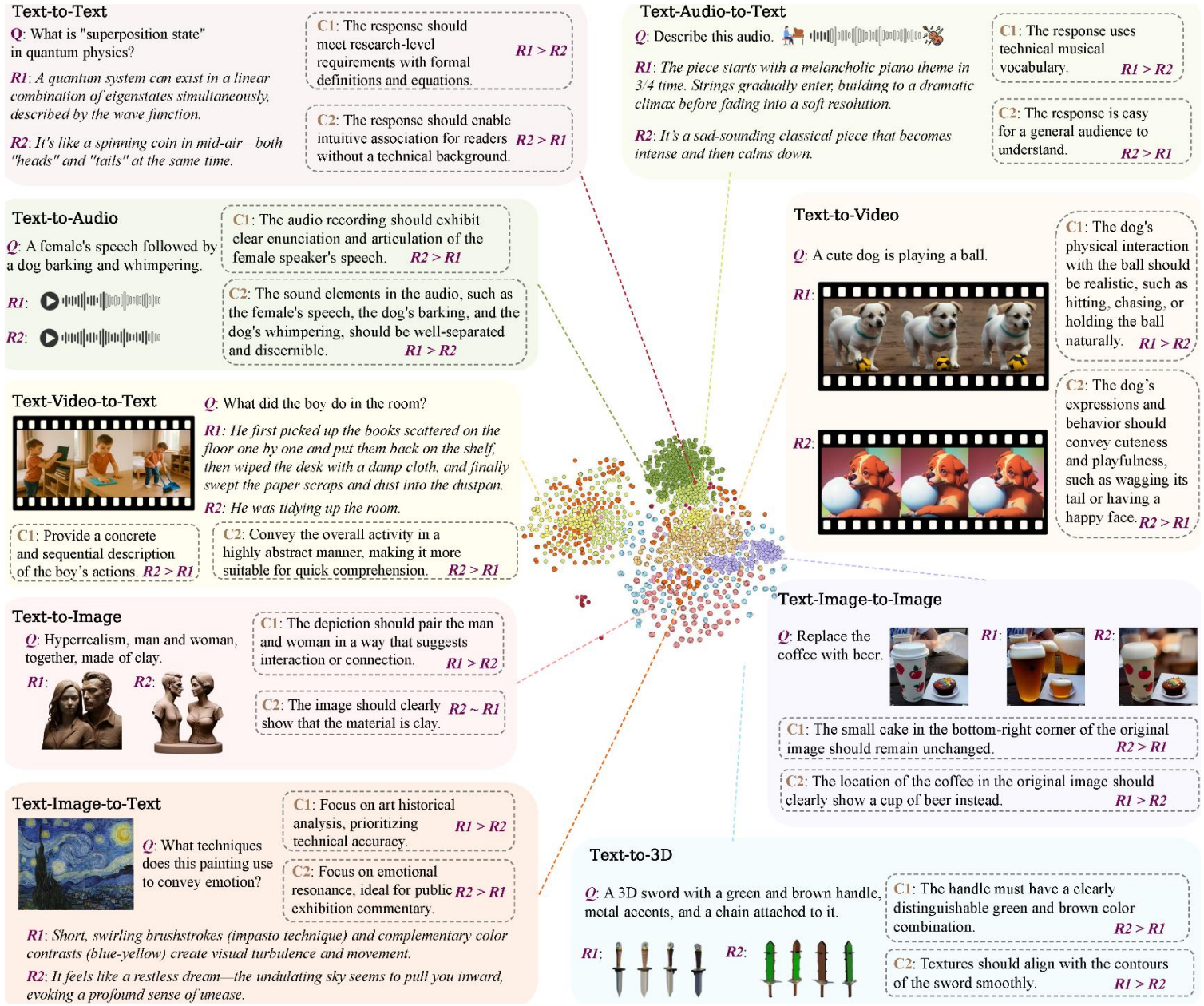
- **Omni-RewardBench**
  - A comprehensive benchmark covering **9 tasks across 5 modalities** with **3,725 high-quality preference pairs** and free-form preference criteria for evaluation
  - https://hf.co/datasets/HongbangYuan/OmniRewardBench

- **Omni-RewardData**
  - A large-scale multimodal dataset with **317K preference pairs**, including **248K general** and **69K instruction-based, free-form preferences**
  - https://hf.co/datasets/jinzhuoran/OmniRewardData

- **Omni-RewardModel**
  - A generalist reward model supporting both **discriminative** and **generative** reward modeling
  - https://hf.co/jinzhuoran/OmniRewardModel

Omni-RewardBench

- Text-to-Text (T2T)
- Text-Image-to-Text (TI2T)
- Text-Video-to-Text (TV2T)
- Text-Audio-to-Text (TA2T)
- Text-to-Image (T2I)
- Text-to-Video (T2V)
- Text-to-Audio (T2A)
- Text-to-3D (T23D)
- Text-Image-to-Image (TI2I)



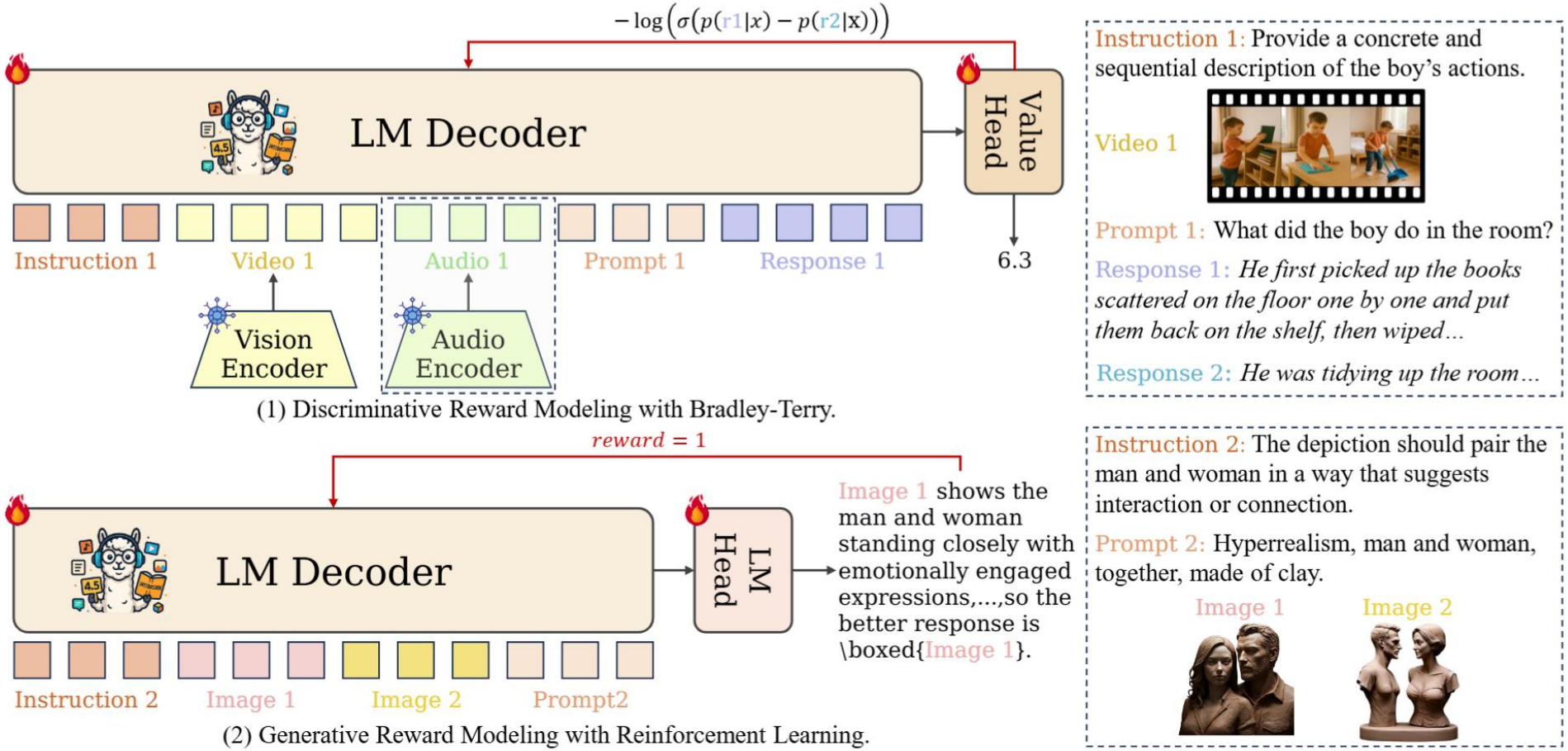
Omni-RewardData

- **248K general preference pairs** (shared human values)
- **69K instruction-based, free-form preference pairs** (user-specific criteria)
- **T2T, TI2T, T2I, and T2V**

Task	Subset	#Size
T2T	Skywork-Reward-Preference	50,000
	Omni-Skywork-Reward-Preference*	16,376
	Omni-UltraFeedback*	7,901
TI2T	RLAIF-V	83,124
	OmniAlign-V-DPO	50,000
	Omni-RLAIF-V*	15,867
	Omni-VLFeedback*	12,311
T2I	HPDv2	50,000
	EvalMuse	2,944
	Omni-HPDv2*	8,959
	Omni-Open-Image-Preferences*	8,105
T2V	VideoDPO	10,000
	VisionRewardDB-Video	1,795

Omni-RewardModel

- **Discriminative RM (BT):** Directly assigns scalar reward scores using the Bradley–Terry objective for efficient preference prediction
- **Generative RM (R1):** Generates reasoning before predicting preferences, enabling interpretable and flexible reward modeling



Experimental Results on Omni-RewardBench

- *Under the w/ Tie setting*

Model	T2T	TI2T	TV2T	TA2T	T2I	T2V	T2A	T23D	TI2I	Overall
Open-Source Models										
Phi-4-Multimodal-Instruct	70.98	53.60	62.53	55.74	35.36	32.14	44.77	24.17	22.71	44.67
Qwen2.5-Omni-7B	65.71	55.11	56.66	59.66	55.99	50.85	32.60	43.71	43.23	51.50
MiniCPM-o-2.6	61.39	51.89	60.95	60.50	47.35	39.70	21.90	37.09	39.30	46.67
MiniCPM-V-2.6	57.55	54.73	53.27	-	48.92	44.61	-	39.40	36.68	47.88
LLaVA-OneVision-7B-ov	50.84	42.23	45.37	-	43.42	40.08	-	35.43	37.12	42.07
Mistral-Small-3.1-24B-Instruct-2503	74.58	57.98	68.62	-	58.55	59.92	-	60.60	62.88	63.30
Skywork-R1V-38B	77.94	59.47	67.72	-	47.94	45.94	-	43.71	41.92	54.95
Qwen2-VL-7B-Instruct	63.55	55.30	59.37	-	33.20	61.25	-	42.38	10.04	46.44
Qwen2.5-VL-3B-Instruct	53.00	49.05	51.24	-	47.74	51.23	-	45.36	44.54	48.88
Qwen2.5-VL-7B-Instruct	68.59	53.03	68.40	-	60.51	47.83	-	50.99	41.05	55.77
Qwen2.5-VL-32B-Instruct	74.82	60.23	63.88	-	60.51	62.38	-	62.58	69.43	64.83
Qwen2.5-VL-72B-Instruct	76.98	61.17	68.40	-	58.94	56.52	-	59.60	62.01	63.37
InternVL2_5-4B	57.55	50.76	55.30	-	48.72	47.07	-	47.35	47.16	50.56
InternVL2_5-8B	60.43	49.62	54.63	-	54.42	49.53	-	42.72	44.10	50.78
InternVL2_5-26B	64.75	57.01	62.98	-	56.97	49.72	-	57.28	48.03	56.68
InternVL2_5-38B	69.06	54.73	64.56	-	54.81	40.26	-	55.96	46.72	55.16
InternVL2_5-8B-MPO	65.95	52.46	68.17	-	56.97	52.55	-	52.98	41.05	55.73
InternVL2_5-26B-MPO	70.74	60.98	**70.43**	-	58.74	47.26	-	56.95	48.03	59.02
InternVL3-8B	76.02	58.71	67.95	-	57.37	48.77	-	51.66	43.67	57.74
InternVL3-9B	73.86	57.39	66.59	-	57.37	51.80	-	60.93	47.16	59.30
InternVL3-14B	76.74	61.74	68.62	-	60.51	61.25	-	59.27	55.02	63.31
Gemma-3-4B-it	74.34	56.82	68.40	-	60.31	60.30	-	54.64	54.15	61.28
Gemma-3-12B-it	73.62	58.52	66.14	-	59.33	62.57	-	56.95	56.33	61.92
Gemma-3-27B-it	77.22	61.17	67.04	-	59.14	61.44	-	63.91	65.94	65.12
Proprietary Models										
GPT-4o	**78.18**	61.74	69.30	62.75	59.33	65.03	44.53	**70.86**	**69.87**	64.62
Gemini-1.5-Flash	72.90	58.52	68.62	57.42	62.48	63.52	32.85	62.25	63.32	60.21
Gemini-2.0-Flash	74.10	54.92	60.50	61.90	62.28	67.49	31.87	68.54	65.50	60.79
GPT-4o-mini	76.50	60.23	67.95	-	57.56	65.22	-	60.26	60.26	64.00
Claude-3-5-Sonnet-20241022	76.74	61.55	67.04	-	61.69	64.27	-	68.54	65.94	**66.54**
Claude-3-7-Sonnet-20250219-Thinking	75.78	**63.83**	68.85	-	62.28	62.38	-	68.21	63.76	66.44
Specialized Models										
PickScore	42.93	43.56	46.95	-	60.12	66.92	-	59.27	51.53	53.04
HPSv2	43.41	45.27	44.70	-	**63.85**	64.65	-	61.26	55.02	54.02
InternLM-XComposer2.5-7B-Reward	59.95	52.65	65.69	-	45.19	61.25	-	43.05	9.61	48.20
UnifiedReward	60.19	53.22	69.53	-	59.72	**70.32**	-	59.93	42.36	59.32
UnifiedReward1.5	59.47	54.17	69.30	-	58.35	69.57	-	61.59	45.41	59.69
`Omni-RewardModel-R1`	71.22	56.06	63.88	-	61.69	58.22	-	63.91	46.29	60.18
`Omni-RewardModel-BT`	75.30	60.23	68.85	**70.59**	58.35	64.08	**63.99**	67.88	58.95	65.36
Average	67.32	55.52	63.02	59.66	55.31	55.59	34.75	53.98	48.60	56.68

- *Under the w/o Tie setting*

Model	T2T	TI2T	TV2T	TA2T	T2I	T2V	T2A	T23D	TI2I	Overall
Open-Source Models										
Phi-4-Multimodal-Instruct	81.15	68.14	74.74	63.47	46.03	51.72	55.05	39.02	49.28	58.73
Qwen2.5-Omni-7B	82.79	68.14	78.16	63.77	65.53	63.09	50.76	56.44	54.11	64.75
MiniCPM-o-2.6	74.04	66.05	71.58	69.76	58.50	61.16	54.80	54.92	48.79	62.18
MiniCPM-V-2.6	74.86	65.12	69.47	-	57.37	58.15	-	51.14	53.62	61.39
LLaVA-OneVision-7B-ov	66.67	57.67	53.42	-	51.93	51.72	-	43.94	43.48	52.69
Mistral-Small-3.1-24B-Instruct-2503	84.43	65.79	79.47	-	65.99	68.67	-	67.80	71.98	72.02
Skywork-R1V-38B	**88.25**	74.42	76.84	-	55.10	57.94	-	45.83	52.66	64.43
Qwen2-VL-7B-Instruct	79.78	70.00	76.58	-	37.41	68.03	-	47.35	12.08	55.89
Qwen2.5-VL-3B-Instruct	68.58	66.05	60.00	-	52.15	60.09	-	51.89	53.62	58.91
Qwen2.5-VL-7B-Instruct	80.87	66.28	78.95	-	65.53	64.59	-	64.77	50.72	67.39
Qwen2.5-VL-32B-Instruct	86.34	74.19	77.37	-	70.29	70.39	-	68.56	70.05	73.88
Qwen2.5-VL-72B-Instruct	87.70	74.65	**80.53**	-	71.88	67.17	-	66.67	69.57	74.02
InternVL2_5-4B	69.95	63.49	64.47	-	58.50	54.94	-	50.38	41.55	57.61
InternVL2_5-8B	72.13	64.88	65.00	-	64.40	61.59	-	58.33	53.14	62.78
InternVL2_5-26B	77.60	72.79	76.32	-	68.03	62.88	-	68.56	59.90	69.44
InternVL2_5-38B	84.15	66.05	70.53	-	66.67	63.30	-	68.94	57.97	68.23
InternVL2_5-8B-MPO	75.96	65.12	77.63	-	65.99	61.80	-	62.88	55.07	66.35
InternVL2_5-26B-MPO	80.87	73.72	**80.53**	-	68.93	62.66	-	67.80	60.87	70.77
InternVL3-8B	84.70	71.63	76.84	-	69.39	65.67	-	59.85	53.62	68.81
InternVL3-9B	83.06	70.23	78.42	-	65.31	65.67	-	71.97	58.45	70.44
InternVL3-14B	85.79	74.65	77.11	-	72.79	68.24	-	68.56	58.94	72.30
Gemma-3-4B-it	83.88	73.02	77.37	-	72.34	66.09	-	67.05	63.77	71.93
Gemma-3-12B-it	81.69	72.09	78.42	-	71.20	71.03	-	67.05	65.70	72.45
Gemma-3-27B-it	**88.25**	75.58	78.16	-	68.48	71.03	-	73.86	71.50	75.27
Proprietary Models										
GPT-4o	86.89	75.58	77.11	70.96	69.61	73.18	53.28	77.65	**73.91**	73.13
Gemini-1.5-Flash	83.88	69.53	78.16	62.28	71.43	71.89	40.66	74.24	73.43	69.50
Gemini-2.0-Flash	85.25	67.91	75.26	67.96	70.52	74.25	60.86	**79.17**	71.98	72.57
GPT-4o-mini	87.43	74.65	77.89	-	67.80	74.89	-	71.59	66.67	74.42
Claude-3-5-Sonnet-20241022	**88.25**	**76.28**	78.68	-	70.75	72.53	-	77.65	72.46	**76.66**
Claude-3-7-Sonnet-20250219-Thinking	84.43	**76.28**	77.89	-	70.07	70.60	-	76.89	72.46	75.52
Specialized Models										
PickScore	49.18	53.49	54.47	-	69.61	75.97	-	67.05	57.49	61.04
HPSv2	49.18	55.12	51.58	-	**73.70**	73.61	-	70.45	60.87	62.07
InternLM-XComposer2.5-7B-Reward	68.85	64.19	74.74	-	51.47	68.24	-	46.59	56.04	61.45
UnifiedReward	68.58	59.77	79.47	-	68.93	**79.83**	-	68.56	46.86	67.43
UnifiedReward1.5	67.76	67.39	78.68	-	67.57	78.97	-	70.45	50.72	68.79
`Omni-RewardModel-R1`	81.77	69.53	75.53	-	71.20	62.02	-	72.35	55.56	69.71
`Omni-RewardModel-BT`	85.79	72.79	79.47	**75.45**	67.12	72.75	**66.41**	77.65	65.70	73.68
Average	78.38	68.57	73.77	66.37	64.61	66.62	52.57	63.54	58.10	67.29

Experimental Results on VL-RewardBench

Models	General	Hallucination	Reasoning	Overall Acc	Macro Acc
Open-Source Models					
LLaVA-OneVision-7B-ov	32.2	20.1	57.1	29.6	36.5
Molmo-7B	31.1	31.8	56.2	37.5	39.7
InternVL2-8B	35.6	41.1	59.0	44.5	45.2
Llama-3.2-11B	33.3	38.4	56.6	42.9	42.8
Pixtral-12B	35.6	25.9	59.9	35.8	40.4
Molmo-72B	33.9	42.3	54.9	44.1	43.7
Qwen2-VL-72B	38.1	32.8	58.0	39.5	43.0
NVLM-D-72B	38.9	31.6	62.0	40.1	44.1
Llama-3.2-90B	42.6	57.3	61.7	56.2	53.9
Proprietary Models					
Gemini-1.5-Flash	47.8	59.6	58.4	57.6	55.3
Gemini-1.5-Pro	50.8	72.5	64.2	67.2	62.5
Claude-3.5-Sonnet	43.4	55.0	62.3	55.3	53.6
GPT-4o-mini	41.7	34.5	58.2	41.5	44.8
GPT-4o	49.1	67.6	**70.5**	65.8	62.4
Specialized Models					
LLaVA-Critic-8B	54.6	38.3	59.1	41.2	44.0
IXC-2.5-Reward	**84.7**	62.5	62.9	65.8	**70.0**
UnifiedReward	60.6	78.4	60.5	66.1	66.5
Skywork-VL-Reward	66.0	80.0	61.0	73.1	69.0
`Omni-RewardModel-BT`	50.8	**89.3**	60.4	**76.3**	66.8

Impact of Training Data Composition

- **Mixed Multimodal Data**
  - Training on single modalities → marginal improvements
  - Training on mixed multimodal data → significantly better generalization

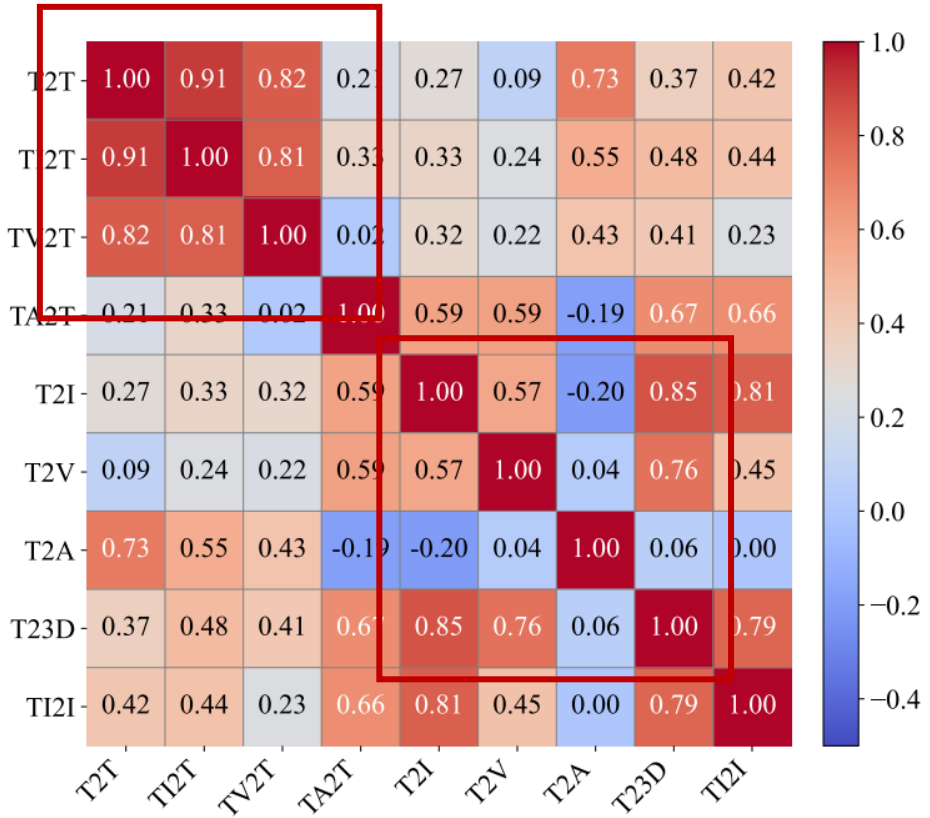
- **Instruction-Tuning Data**
  - Removing instruction-based data → clear performance drop
  - Instruction tuning is critical for fine-grained reward modeling

Model	T2T	TI2T	TV2T	TA2T	T2I	T2V	T2A	T23D	TI2I	Overall
MiniCPM-o-2.6	61.39	51.89	60.95	60.50	47.35	39.70	21.90	37.09	39.30	46.67
w/ T2T	74.30	54.73	66.37	69.75	45.38	43.86	55.96	49.67	54.15	57.13
w/ TI2T	74.54	59.62	66.82	69.75	41.45	48.77	61.31	51.00	56.33	58.84
w/ T2I & T2V	52.28	45.83	51.47	59.38	**58.93**	**64.84**	56.93	67.55	**60.26**	57.50
w/ Full	**75.30**	**60.23**	**68.85**	**70.59**	58.35	64.08	**63.99**	**67.88**	58.95	**65.36**
w/ Preference-Only	54.92	49.80	64.79	55.74	59.14	61.06	64.00	64.90	53.71	58.67

Correlation of Performance Across Tasks

- **Understanding Tasks** (text, image, video understanding)
  - High correlation: 0.8 – 0.9

- **Generation Tasks** (image, video, 3D generation)
  - Moderate to high correlation: 0.7 – 0.8



Future Work

- **Scaling to Richer Modalities**
  - Extend reward modeling to richer and more complex modalities, such as embodied agents and GUI-based agents, where evaluation involves interactive behaviors and long-horizon decision-making

- **Agentic Reward System**
  - Integrate multiple specialized reward experts, each focusing on different aspects such as modality, task, or preference type, with dynamic selection and aggregation for adaptive evaluation

- **Heterogeneous Preference Alignment**
  - Move beyond binary preference learning to model heterogeneous and multi-dimensional human preferences across diverse users, contexts, and scenarios