SoliReward: Mitigating Susceptibility to Reward Hacking and Annotation Noise in Video Generation Reward Models

Supplementary Material

6. More Discussions about BT-WT

While 'pass' can form win-ties in single-item annotation, the applicability is dimension-dependent. Binary annotation inherently compresses a continuous quality spectrum into two discrete points. For some dimensions, 'pass' denotes an absolute, discrete state. While for others, 'pass' merely signifies surpassing a subjective threshold, masking underlying gradations of quality. We propose a test to assess a dimension's suitability:

> Does a shared 'pass' label for samples y_i and y_j imply identical underlying quality, or merely a negligible difference?

Yes: The dimension is suitable for constructing win ties. **No** (i.e., y_i may still be superior to y_j): The dimension is unsuitable for win ties.

7. Additonal Experiments

7.1. Pairing Strategy

We ablate the impact of data pairing strategies on reward model performance by comparing two approaches: the in-prompt and cross-prompt pairing strategies. In the in-prompt method, both videos within a preference pair are generated from an identical prompt. Conversely, the cross-prompt strategy permits a pair to be formed from videos generated by different prompts. We construct separate training datasets using each strategy and evaluated the resulting RMs' accuracy and reward margin on fixed evaluation datasets. As shown in Tab. 7 and Tab. 8, the cross-prompt strategy achieves performance comparable to its in-prompt counterpart. Furthermore, a hybrid strategy combining both datasets also yielded similar results. The primary advantage of the cross-prompt approach is its relaxed data requirement: while the in-prompt method necessitates two or more video generations per prompt, the cross-prompt strategy can effectively leverage data from prompts that yielded only a single video. This improves the utilization of available data without compromising reward model performance.

7.2. Model Scaling

As demonstrated in Tab. 9, our analysis of model scaling reveals two key findings. First, performance improvements are substantially more pronounced in the OOD evaluation than in the ID setting. Second, these scaling benefits are not monotonic and are heavily concentrated in the transition from 1B to 8B parameters. This initial scaling step yields significant OOD accuracy gains (up to +5.01) and dramatically improves the OOD reward margin (e.g., from 2.95 to 5.51 for "Phy & Deform"), whereas ID accuracy sees only modest increases. In contrast, scaling further from 8B to 14B yields diminishing returns.

The observed diminishing returns when scaling from 8B to 14B parameters, following significant gains from 1B to 8B, suggest a multi-faceted bottleneck. The initial 1B model is likely under-parameterized, lacking the capacity to capture essential features, which the 8B model successfully acquires. However, the 8B model may already achieve

Table 7. Impact of pairing-strategy for reward model training. Accuracy is evaluated on both ID and OOD datasets. The cross-prompt strategy is comparable with in-prompt strategy.

Task	Approach	Reward ACC	
		ID	OOD
Phy & Deform	Cross-Prompt	76.74	79.54
	In-Prompt	76.77	79.22
	Hybrid	76.09	79.16
TA	Cross-Prompt	76.39	60.25
	In-Prompt	76.67	59.26
	Hybrid	75.64	59.41

Table 8. Impact of pairing-strategy for reward model training. Pair score margin is evaluated on both ID and OOD datasets. The cross-prompt strategy is comparable with in-prompt strategy.

Task	Approach	Reward Margin	
		ID	OOD
Phy & Deform	Cross-Prompt	3.83	4.17
	In-Prompt	3.98	3.77
	Hybrid	3.98	4.13
TA	Cross-Prompt	2.93	1.47
	In-Prompt	2.73	1.28
	Hybrid	2.87	1.37

Table 9. Influence of model size on reward model accuracy and reward margin between positive and negative samples.

Task	Approach	ID		OOD	
		ACC	Margin	ACC	Margin
Phy & Deform	InternVL3-1B	76.25	3.10	77.65	2.95
	InternVL3-8B	78.48	3.60	81.43	5.51
	InternVL3-14B	78.57	3.69	81.71	5.19
TA	InternVL3-1B	77.48	2.08	58.45	0.33
	InternVL3-8B	79.02	2.11	63.46	0.92
	InternVL3-14B	78.83	2.15	65.29	0.71

Table 10. Comparison of reward model performance trained via BT, BTT and BT-WT. Reward model accuracy and VBench2 Human Fidelity scores are reported.

Method	Reward Model	Post-Training	
	ACC	VBench2	MQ
BT	77.63	0.8693	0.1719
BTT	77.78	0.8700	0.0690
BT-WT	**78.27**	**0.8999**	**0.3302**

capacity saturation for the given task's intrinsic complexity. This means the additional parameters of the 14B model offer only marginal utility. Furthermore, performance is likely becoming data-limited, where the 14B model requires a larger or more diverse dataset than available to unlock further gains. This is supported by the degradation in the OOD reward margin for both tasks, which indicates the 14B model may be overfitting to the training data, a common challenge when model size outpaces data scale and optimization refinement.

7.3. BT, BTT and BT-WT

We evaluate the impact of the BT, Bradley-Terry with Ties (BTT), and BT-WT loss functions on RM training and subsequent policy optimization. The three losses utilize different data pairing strategies: BT uses only win-lose pairs; BTT incorporates win-lose, win-tie, and lose-tie pairs; and BT-WT employs win-lose and win-tie pairs. The BTT data consists of the BT-WT data plus an additional $150k$ lose-tie pairs.

As discussed in Sec. 3.3, VideoAlign [17] retains all tie samples, using a loss function that models A-wins, B-wins, or ties (which includes both win-tie and lose-tie pairs). We contend this approach is ill-suited for our annotation scenario. Specifically, if two negative samples are labeled independently (e.g., for deformity), we cannot assume their degree of severity is equivalent. Erroneously treating them as a 'tie' (a lose-tie pair) would introduce label noise, diminishing the RM's discriminative capacity and ultimately degrading policy performance.

Empirical results in Tab. 10 support this argument: BTT achieves a lower RM accuracy than BT-WT. Furthermore, the degradation extends to generation quality. As shown in the post-training metrics, the BTT approach results in inferior VBench2 Human Fidelity scores compared to BT-WT. These findings confirm that incorporating lose-tie pairs into the loss function is suboptimal for our data distribution.

7.4. HPQA Layer Indices

Empirical results in Table 11 and Table 12 demonstrate that the HPQA architecture maintains its efficacy across diverse VLM backbones. Furthermore, our ablation study indicates that aggregating representations from 4 to 5 intermediate layers yields better performance.

Table 11. Performance of Qwen2-VL (2B) [31] across different layer indices.

Task	Layer Indices	ID		OOD	
		ACC	Margin	ACC	Margin
Phy & Deform	'Yes' Token	75.40	1.71	77.05	1.69
	{14, 28}	75.79	1.71	77.24	1.85
	{9, 18, 28}	75.20	1.64	77.59	1.59
	{7, 14, 21, 28}	**76.34**	1.72	77.78	1.81
	{6, 12, 18, 24, 28}	76.29	1.77	**78.89**	1.84
	{5, 10, 14, 19, 23, 28}	74.11	1.67	75.73	1.56
TA	'Yes' Token	77.33	3.41	55.28	0.70
	{7, 14, 21, 28}	**78.62**	3.55	56.19	0.56
	{5, 10, 14, 19, 23, 28}	77.53	3.17	**56.33**	0.54

Table 12. Performance of Qwen2.5-VL (3B) and InternVL3 (14B) across different layer indices.

Task	Model	Layer Indices	ID		OOD	
			ACC	Margin	ACC	Margin
Phy & Deform	Qwen2.5-VL (3B)	'Yes' Token	**75.45**	1.61	40.76*	-1.52*
		{9, 18, 27, 36}	73.68	1.61	76.38	1.50
		{7, 14, 21, 28, 36}	73.51	1.47	**76.61**	1.39
		{6, 12, 18, 24, 30, 36}	72.87	1.43	74.27	1.35
	InternVL3 (14B)	{16, 32, 48}	78.10	3.56	80.99	5.39
		{12, 24, 36, 48}	**78.50**	3.52	80.31	5.17
		{9, 18, 27, 36, 48}	78.20	3.71	**81.16**	5.21
TA	Qwen2.5-VL (3B)	'Yes' Token	76.19	1.89	27.73*	0.24*
		{9, 18, 27, 36}	76.54	2.84	**59.06**	0.51
		{6, 12, 18, 24, 30, 36}	**77.23**	2.80	58.16	0.35

* scores degenerate to discrete values.

Table 13. Comparison of different approaches on human preference benchmarks from VBench2 [41]. The evaluation metrics include Anatomy (Human Anatomy), Clothes (Human Clothes), Identity (Human Identity), and Fidelity (Human Fidelity).

Method	Anatomy	Clothes	Identity	Fidelity (Overall)
Baseline	0.8915	0.8905	0.7457	0.8426
VideoAlign-MQ	0.9009	0.8714	0.8361	0.8695
'Yes' Token w/ BT	0.9154	0.9067	0.7778	0.8666
'Yes' Token w/ BT-WT	**0.9312**	**0.9259**	0.8013	0.8861
HPQA w/ BT	0.9153	0.9064	0.7861	0.8693
HPQA w/ BTT	0.9254	0.8922	0.7923	0.8700
HPQA w/ BT-WT	0.9164	0.9231	**0.8601**	**0.8999**

7.5. Post-Training

7.5.1. Physical Plausibility and Subject Deformity

We select corresponding VBench2 Human Fidelity, including human anatomy, human clothes and human identity, as the evaluation dimensions. As detailed in Tab. 13, our RM outperforms both the baseline and the VideoAlign-MQ guided training. Both BT-WT and HPQA generate positive returns, with the former yielding even higher returns.

7.5.2. Semantic Alignment

Tab. 14 summarizes the performance comparison of the HunyuanVideo backbone after post-training with different reward models, evaluated on the VBench benchmark for semantic alignment. Both post-training methods, using the

VideoAlign TA reward model and our SoliReward, improve upon the baseline HunyuanVideo model's semantic score of 0.7334. Our proposed reward model achieves the highest overall performance, with a semantic score of **0.7544**, surpassing the TA model's score of 0.7421.

8. Data Annotation

8.1. Single-Item Binary Annotation Design

To ensure a standardized and rigorous evaluation process, we formulate detailed annotation guidelines for our human annotators. These guidelines, summarized in Tab. 15, provide a structured framework for assessing the quality of generated videos. The evaluation is organized into three primary dimensions. The first dimension, **Subject Deformity**, focuses on the structural plausibility and temporal stability of the subjects within the video. The second dimension, **Physical Plausibility**, evaluates the adherence of the video's dynamics to real-world physical laws, including motion, gravity, and object interactions. The final dimension, **Semantic Alignment**, measures the relevance of the generated video content to the input text prompt, encompassing core semantics, detailed descriptions, and stylistic specifications. To simplify the annotation task and ensure consistency, annotators are instructed to perform a binary assessment for each dimension. They are required solely to judge whether a video passes or fails the specific criteria defined for that dimension.

8.2. Data Distribution

We collected annotations for 250k in-house videos and a 50k-sample out-of-distribution (OOD) test set generated by other SOTA models (including Wan2.1 [30], Wan2.2 [30], Veo 3 [7] and Seedance 1.0 [6]). These samples originate from 20k unique prompts. To ensure diversity, we balanced these prompts across multiple attributes including subject, motion, style, and camera control. In Fig. 7, we demonstrate the quality distribution of our in-house dataset. To construct the BT-WT dataset, we randomly selected $350k$ win-lose and $150k$ win-tie pairs based on the cross-prompt pairing strategy.

9. Implementation Details

9.1. Reward Model Training

We trained the reward model on a 4-node CentOS cluster, each node featuring 8 NVIDIA H20 GPUs with 96GB CUDA memory and an AMD EPYC 9K84 96-Core Processor. All other hyperparameters are detailed in Tab. 16.

9.2. Post-Training

We conducted post-training experiments on a 5-node CentOS cluster, each node featuring 8 NVIDIA H800 GPUs (80GB CUDA memory) and an Intel Xeon Platinum 8476C Processor. Key hyperparameters are detailed in Tab. 17.

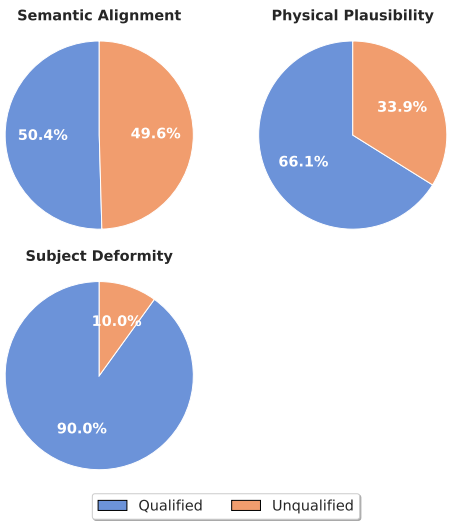

Figure 7. Quality distribution of our in-house dataset.

10. More Visualization Results

Further visualization results are presented in Fig. 8. We also provide visual comparisons between BT and BT-WT in Fig. 9. For extensive qualitative results, please refer to the supplementary material, which includes HTML pages for convenient viewing.

Table 14. Comparison of post-training with different reward models on semantic alignment. Evaluations are conducted on VBench [10].

Backbone	RM	Scene	Consistency	Appearance	Object	Spatial	Action	Temporal	Color	Multiple	Semantic (Avg.)
HunyuanVideo	-	0.3496	0.2700	0.2021	0.8006	0.6993	0.9600	**0.2539**	0.8730	0.6966	0.7334
HunyuanVideo	TA	**0.3895**	**0.2702**	**0.2039**	0.7927	0.7150	**0.9700**	0.2539	0.8269	0.7477	0.7421
HunyuanVideo	Ours	0.3692	0.2678	0.1981	**0.8275**	**0.7517**	**0.9700**	0.2518	**0.9080**	**0.7630**	**0.7544**

Table 15. Definitions of the evaluation dimensions and their key assessment criteria used in our human evaluation framework.

Evaluation Dimension	Definition and Key Criteria
Subject Deformity	Assesses the presence and severity of **structural artifacts and temporal instability** impacting subjects (e.g., humans, animals, objects). - **Structural Artifacts:** Evaluates anatomical incorrectness, penalizing severe distortions, unnatural forms, or implausible subject parts (e.g., faces, limbs). - **Temporal Instability:** Measures inconsistencies in a subject's identity or form, penalizing artifacts like "melting", "flickering", or unnatural morphing across frames.
Physical Plausibility	Assesses the adherence of video dynamics to real-world **physical principles**. - **Motion Dynamics:** Checks if object motion, acceleration, and inertia appear physically coherent and natural. - **Object Interactions:** Evaluates the realism of interactions with forces like gravity (e.g., falling) and between entities (e.g., collisions, splashes). - **Material Dynamics:** Assesses the realistic behavior and deformation of complex materials such as fluids (water, smoke) or soft bodies (cloth).
Semantic Alignment	Assesses the **fidelity** of the generated video content with respect to the input text prompt. - **Core Semantics:** Fidelity to the primary semantic components of the prompt, including the main subject, key action, and overall scene. - **Detailed Attributes:** Fidelity to specific descriptive details, such as subject attributes (e.g., color, appearance), action modifiers, and background elements. - **Stylistic & Cinematic Fidelity:** Alignment with specified artistic styles (e.g., 3D render) and cinematic instructions (e.g., camera motion, "close-up").

Table 16. Implementation details for reward model training.

Hyperparameter	Value
Model	InternVL3-1B/8B/14B
Distributed Training	DeepSpeed Stage 0/3/3
Trainable Parameters	Full
Learning Rate	1e-6
Num. Train Epochs	3.0
Per Device Train Batch Size	1
Gradient Accumulation Steps	10
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.999
Weight Decay	0.01
LR Scheduler	linear
Warmup Ratio	0.05
Reward Margin	3.0
Precision	BF16
Gradient Checkpointing	True

Table 17. Implementation details for post-training.

Hyperparameter	Value
Base Model	HunyuanVideo 14B
Trainable Parameters	Full
Learning Rate	1e-6
Per Device Train Batch Size	1
Gradient Accumulation Steps	4
Max Train Steps	500
LR Scheduler	constant with warmup
LR Warmup Steps	0
Mixed Precision	bf16
Gradient Checkpointing	True
SDE η	0.25
Video Resolution (train)	480×480
Video Frames (train)	32
Video FPS (train)	8
Denoising Steps (train)	16
Time Shift (train)	5.0
Video Resolution (test)	640×640
Video Frames (test)	91
Video FPS (test)	18
Denoising Steps (test)	30
Time Shift (test)	7.0
Group Size	8
Distributed Training	FSDP FULL SHARD

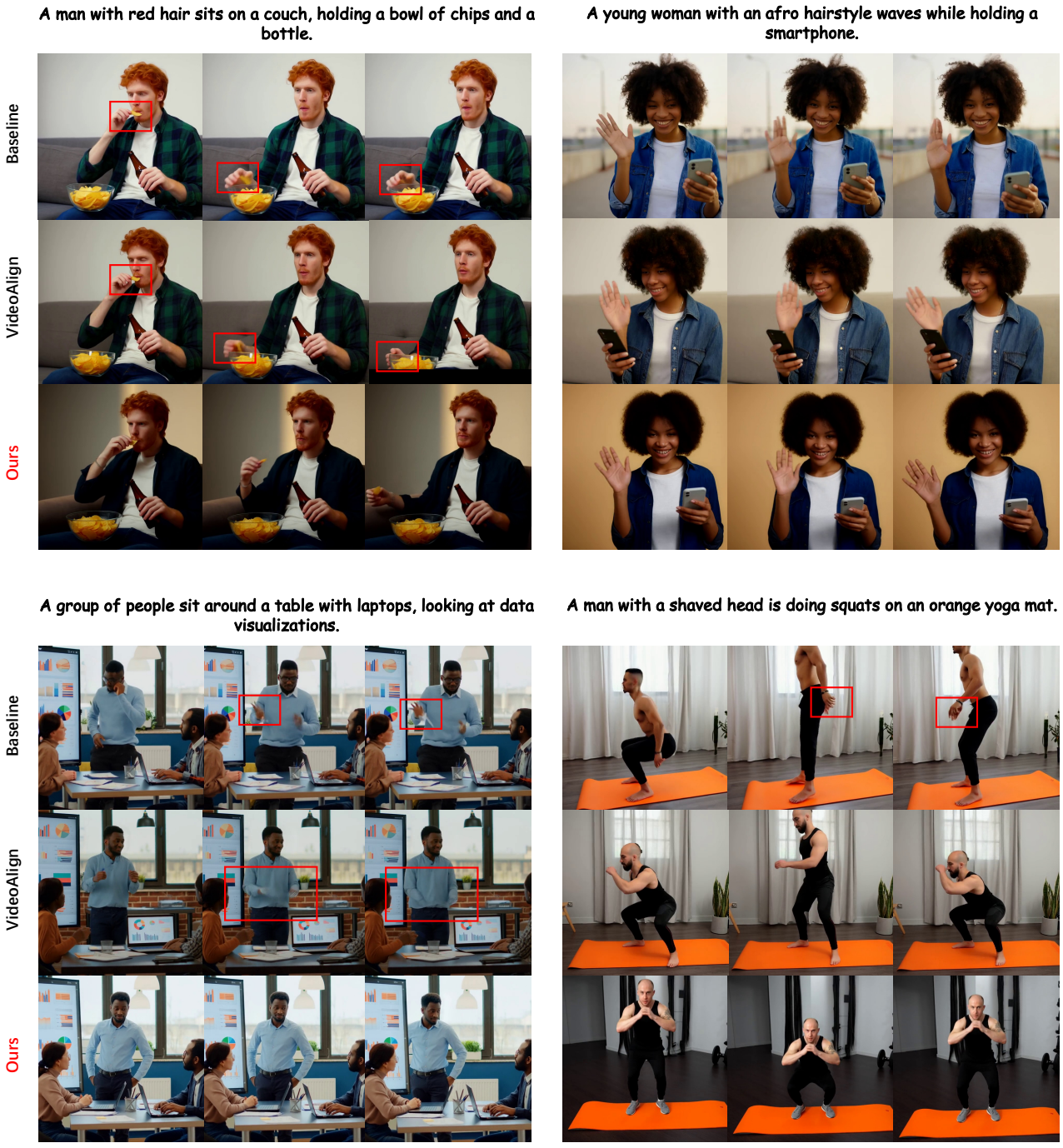

Figure 8. More visualization results guided by VideoAlign and SoliReward.

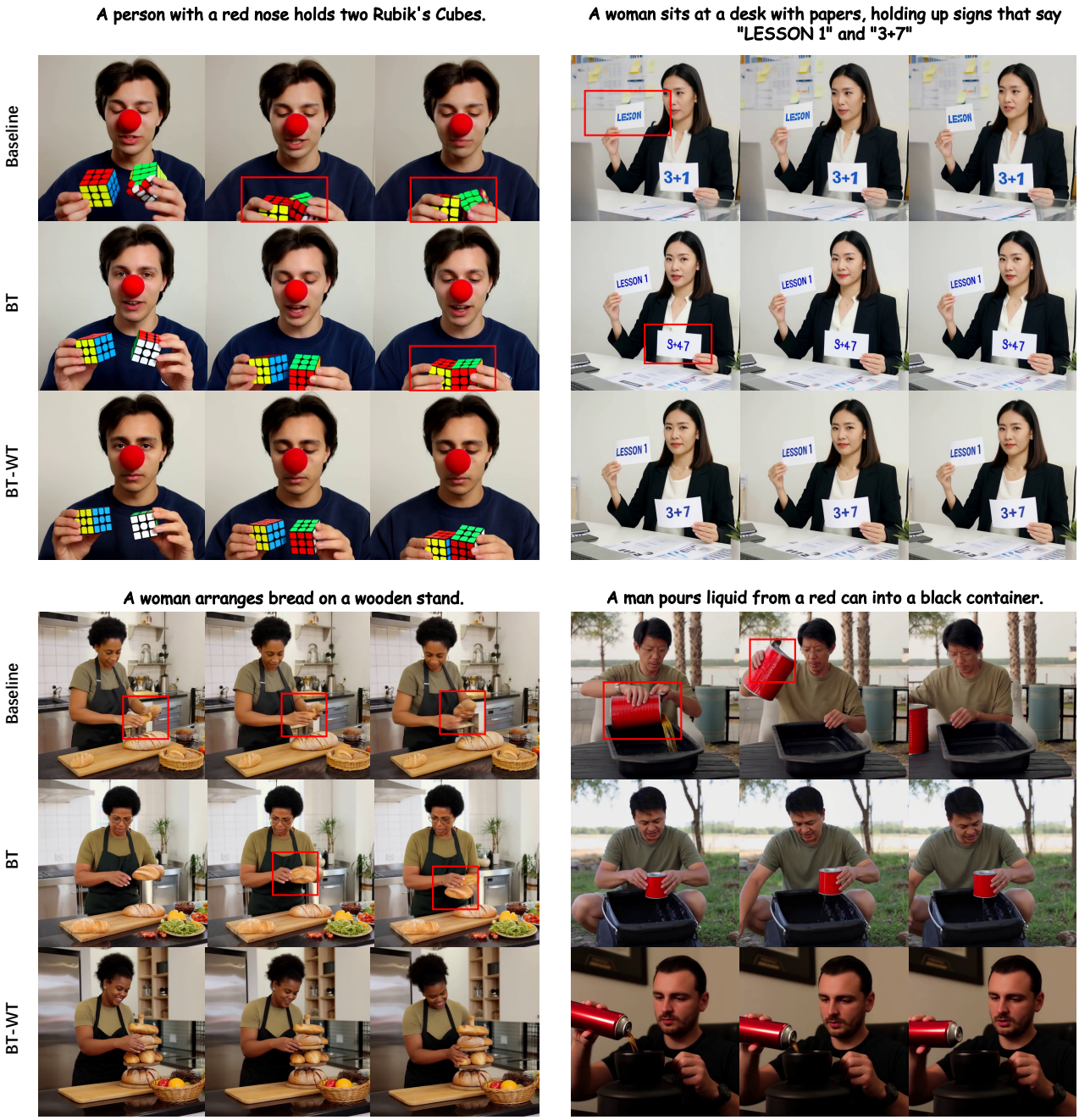

Figure 9. Visualization results guided by BT and BT-WT.