Learning What to Trust: Bayesian Prior-Guided Optimization for Visual Generation

Supplementary Material

6. Additional Experiment Settings

We conduct experiments across three generation tasks with the following configurations. For text-to-video generation, we use a batch size of 32 distributed across 32 H100 GPUs. For image-to-video generation, we employ a batch size of 16 on 16 H100 GPUs. For text-to-image generation, we set the batch size to 8 with 8 H100 GPUs. All models are trained for 200 steps. For video generation tasks, we utilize the iStock dataset, where captions are automatically generated from real videos. Specifically, for image-to-video generation, we apply object removal to the first frame of each real video to increase task difficulty and better evaluate the model's generation capability. The training set comprises 10,000 samples, while the test set contains 100 samples. For text-to-image generation, we adopt the dataset provided by DanceGRPO for training. The evaluation is performed on the HPSv2 benchmark, which consists of 3,200 prompts. All models are trained under BF16 mixed precision using AdamW with fixed denoising steps and rollout samples, and inference uses DDIM sampling with task-specific resolutions.

7. Addtional Results

7.1. Additional Text-to-Video results.

Table 5. Text-to-video results on VBench video-condition consistency metrics (Part 1). We compare our proposed modules against the baseline GRPO. The best results are shown in bold, and the second-best results are underlined. †Our implemented DanceGRPO.

Method	Object Class	Multiple Objects	Human Action	Color
GRPO†	<u>0.6875</u>	0.2111	0.6100	<u>0.8408</u>
RAS	0.4902	0.1021	0.5580	0.8127
CRT	0.5434	<u>0.2485</u>	<u>0.6300</u>	0.8218
RAS+CRT	**0.6899**	**0.2736**	**0.6500**	**0.8594**

Scene	Temporal Style	Overall Consistency	Appearance Style	Spatial Relationship
0.1628	0.2285	0.2179	<u>0.2000</u>	**0.3320**
0.1302	**0.2386**	**0.2306**	**0.2066**	0.2565
0.2278	0.2252	0.2264	0.1912	0.2900
<u>0.2253</u>	<u>0.2342</u>	<u>0.2301</u>	0.1964	<u>0.2907</u>

We evaluated the results using video-condition consistency metrics from VBench [6], which encompasses 9 dimensions. The results are shown in Tab. 5. Our BPGO with the complete two modules (RAS+CTR) achieves the best performance across most dimensions, demonstrating significant improvements over the baseline GRPO. Specifically, our BPGO obtains the highest scores in four metrics. The individual modules also show competitive performance, with RAS excelling in Appearance Style and CTR achieving strong results in Scene. These results validate the effectiveness of our proposed modules, particularly when combined, in enhancing video-condition consistency across diverse evaluation criteria.

Table 6. Human evaluation results on 100 video pairs comparing BPGO against GRPO.

Criterion	**Win**	**Tie**	**Lose**
Overall Quality	34	48	18
Text Alignment	27	56	17

We further conduct a human evaluation on 100 video pairs. As shown in Tab. 6, BPGO is preferred over GRPO in 34% of cases for Overall Quality (48% tied, 18% worse) and in 27% for Text Alignment (56% tied, 17% worse), demonstrating a clear human preference.

7.2. Additional Text-to-Image results.

Tab. 7 shows results of HPSv3 being reward. BPGO continues to outperform the GRPO baseline under HPSv3, while maintaining consistent improvements on pickscore and ImageReward. We also compare BPGO with Flow-GRPO using PickScore as the reward. Superior performance across multiple metrics is shown in Tab. 8

Table 7. Results of using HPSv3 as reward on T2I task.

Method	HPSv3	PickScore	ImageReward
SD-1.4	4.9012	20.6692	0.1058
GRPO†	13.7177	21.5390	0.7957
Ours	13.8293	21.6586	0.9274

7.3. Analysis of Prior Quality and Robustness

To validate that the priors used in BPGO are semantically meaningful rather than arbitrary, we measure their empirical correlation with true rewards during training. Tak-

Table 8. Results of using PickScore as reward on T2I task.

Method	PickScore	HPSv3	ImageReward
FLUX	21.7268	7.1415	1.0056
Flow-GRPO†	21.7192	7.1466	1.0128
Ours	21.7397	7.1632	1.0295

ing the SFT-based prior as a representative case, we observe a statistically significant positive correlation with running rewards (Pearson $r = 0.438$, $p < 0.001$; Spearman $\rho = 0.457$), with closely aligned reward distributions as illustrated in Fig. 8. This confirms that the priors serve as effective reference anchors that reflect task-intrinsic semantic quality.

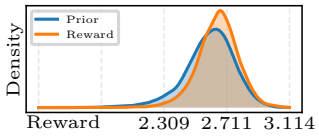


Figure 8. Distributions of running rewards and SFT priors.

To further assess robustness, we introduce synthetic noise into the priors and evaluate the resulting performance. As reported in Tab. 9, BPGO remains stable and maintains competitive performance under noisy prior conditions. This robustness is attributed to the design of BPGO, which incorporates the prior solely through auxiliary regularization without altering the primary GRPO optimization direction, thereby limiting the influence of any biased or misaligned prior signals.

Table 9. Results with noise perturbed into priors. VA = VideoAlign

Noise level	VideoClipXL	VA-TA	VA-overall
No	2.6788	1.1193	-0.0478
20%	2.6590	0.9294	-0.3479
80%	2.6603	0.9052	-0.7993

7.4. Additional Sensitivity Analysis of Hyperparameters

We conduct a joint sensitivity analysis of the two key hyperparameters in BPGO: α, which controls the RAS weighting, and β, which governs the CRT strength.

Consistent with the ablation study in the main paper, training performance is robust to a wide range of α values, with medium values yielding optimal results. Excessively small values under-utilize the prior signal, while excessively large values may over-constrain optimization, yet neither extreme leads to training failure.

As shown in Fig. 9 and Tab. 10, with RAS enabled and $\beta \in \{0.3, 0.7\}$, training remains stable with comparable performance across configurations, demonstrating robustness over a broad parameter range. While larger β values (e.g., $\beta = 0.7$) introduce increased optimization variance, they do not induce rapid training failure. RAS effectively mitigates this variance, maintaining overall training stability.

Table 10. Results under different $alpha$ and $beta$. VA = VideoAlign.

α	β	VideoClipXL	VA-TA	VA-overall
0.5	0.3	2.6804	0.7642	-1.0825
w/o	0.3	2.6675	0.8645	-0.8923
w/o	0.7	2.6683	0.7557	-1.0281

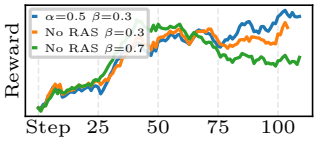


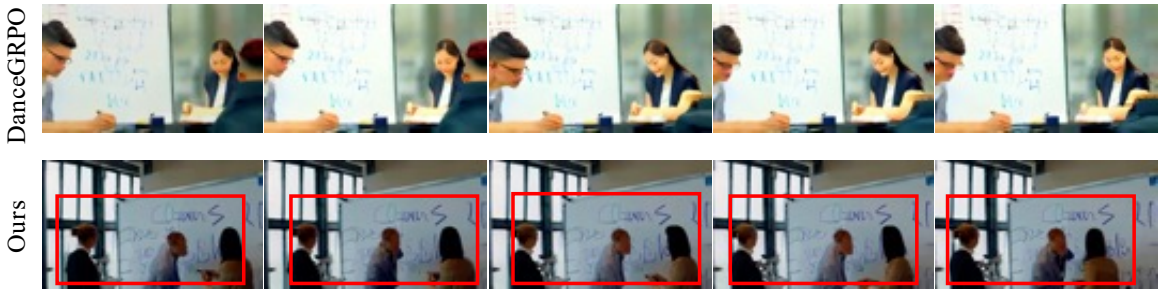
Figure 9. Reward curves under different α and β settings.

7.5. Additional Qualitative Comparison.

Figs. 10 to 13 present qualitative comparisons on text-to-video and image-to-video generation tasks. Our method consistently outperforms DanceGRPO in maintaining consistency with input conditions. For text-to-video generation, our approach accurately captures critical details such as safety equipment, specific actions (e.g., cooking with utensils), object attributes (e.g., cartoonish appearance), and correct object counts (e.g., two children). For image-to-video generation, our method preserves fine-grained visual details while correctly generating specified attributes (e.g., "NEW JOB" text projection), actions (e.g., vaccine administration, makeup application), and multi-subject scenes (e.g., four friends in winter clothing), as highlighted by the red bounding boxes. These examples illustrate our model's superior ability to generate videos that accurately reflect semantic content and fine-grained details from both textual and visual conditions.

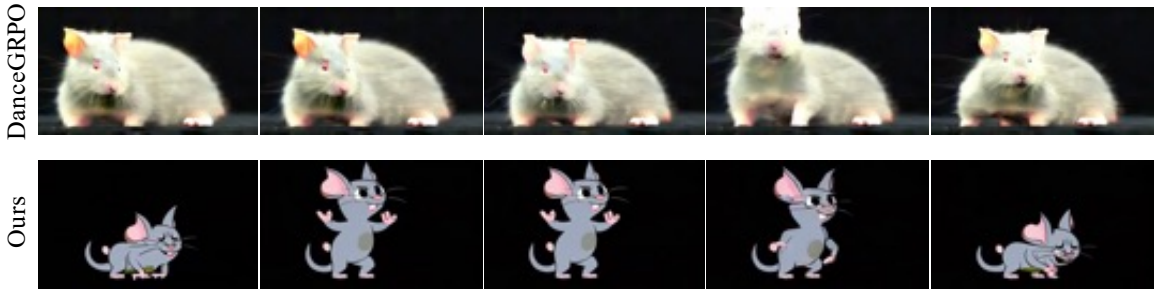
Fig. 14 shows qualitative results on text-to-image generation. Our method (RAS+CRT) outperforms baselines in accurately reflecting textual details. Key improvements include: correct object counts (e.g., "seven people"), fine-grained attributes (e.g., "Tokyo Ghoul mask,"), and specific gestures (e.g., "scissor hand gesture"). While FLUX and DanceGRPO often miss critical details or generate incorrect quantities, our combined approach consistently produces images that faithfully align with the text prompts.

The video depicts a professional setting … The workers, *donning safety gear*, reciprocate with handshake exchanges reflecting mutual respect and teamwork.



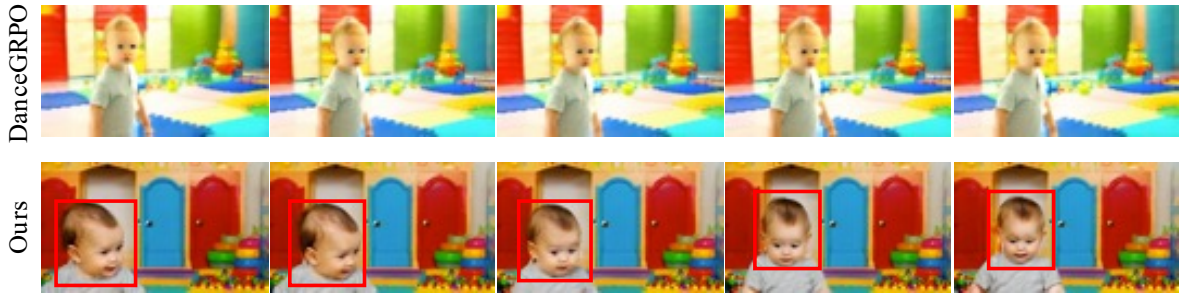
… their *focus directed towards a large whiteboard*. Each person is holding a marker and actively contributing their ideas by *writing or drawing on the surface.* …

… a person is cooking pieces of duck meat in a frying pan. … with the cook *likely using a utensil* to stir the meat, ensuring even cooking. …

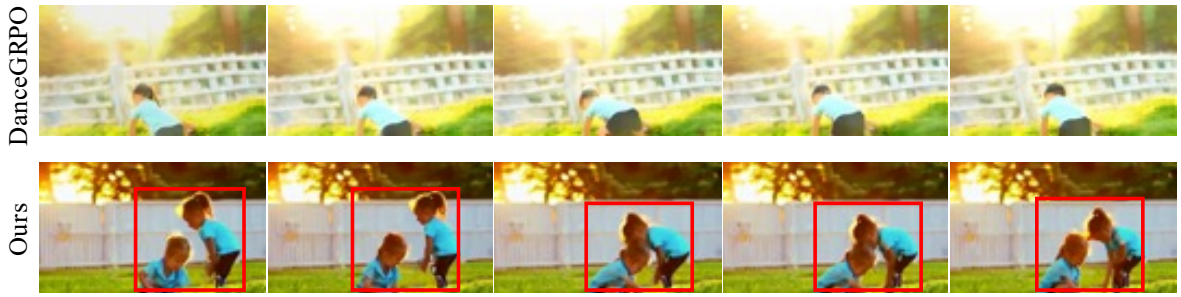


The video features a *cartoonish* grey rat character against a black background. …

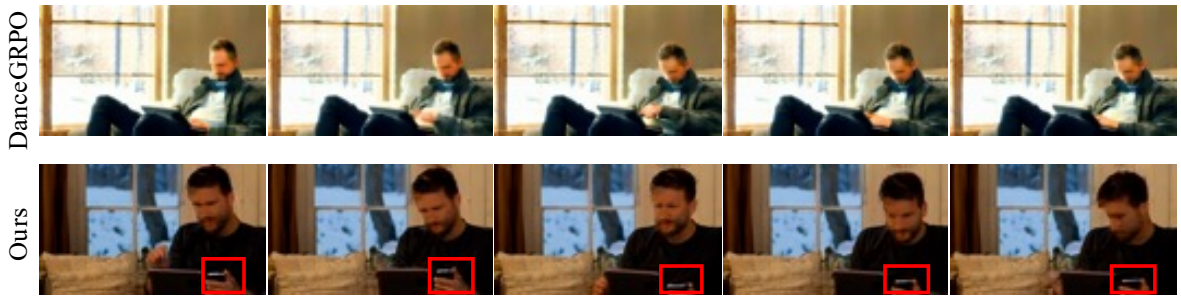
Figure 10. Qualitative comparison on text-to-video task.



A toddler in a gray shirt is standing in …. The camera remains stationary, focusing on the toddler as they *look around with curiosity*, displaying a sense of wonder and exploration. …



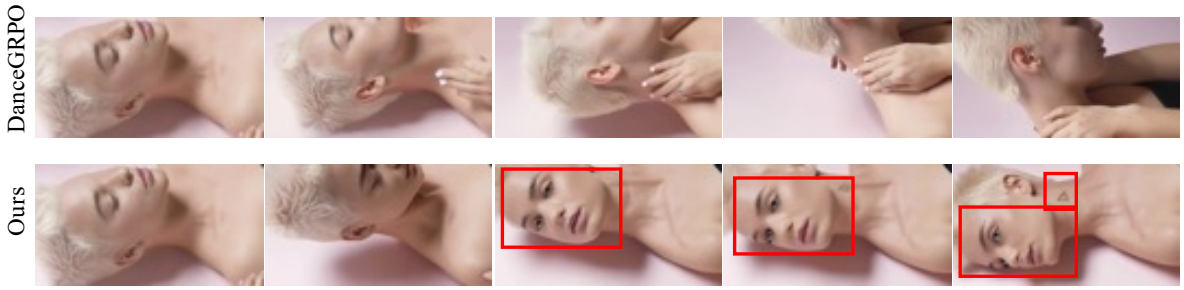
Two children, likely siblings, are engaged in playful interaction with the grass in a sunny outdoor park. …



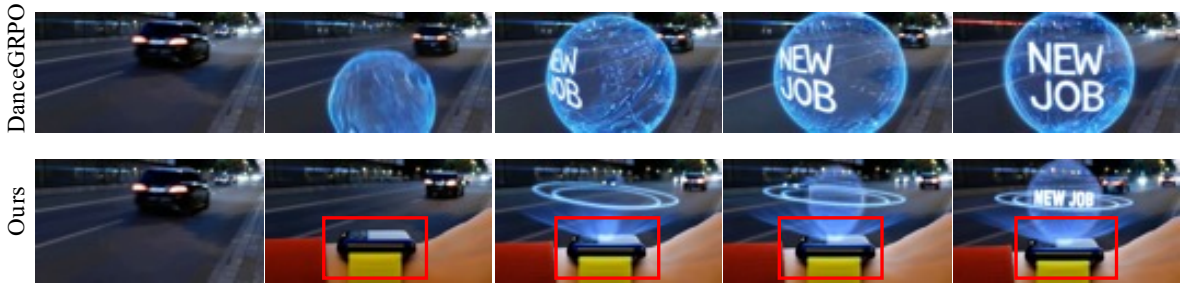
The video features a man seated comfortably on a couch, engaging with both a tablet and *a smartphone*. …

In a modern office setting, *two professionals, a man and a woman*, are engaged in a discussion while standing by large floor-to-ceiling windows. …

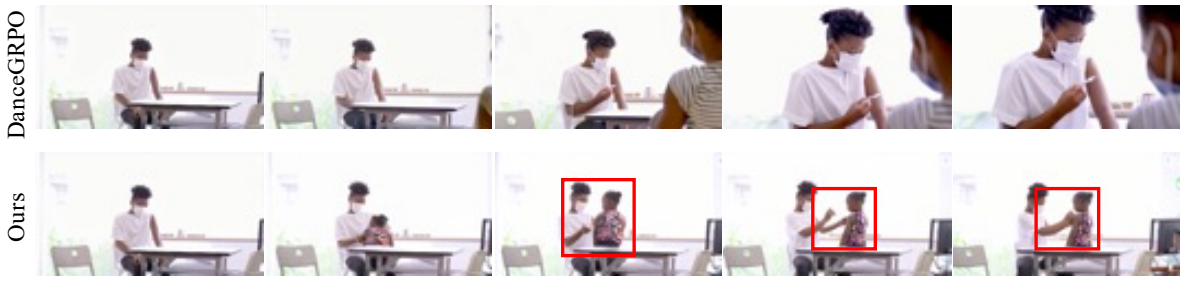
Figure 11. Qualitative comparison on text-to-video task.



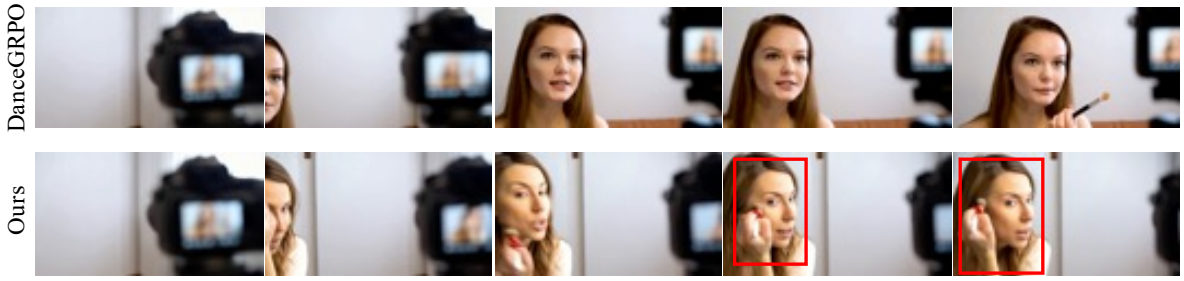
The video features a person with short blonde hair and *a small tattoo of a triangle on their neck*. The individual gently touches their neck with a delicate hand and *then turns their head to gaze off to the side*. …



An urban night scene where *a person's smartwatch is the focal point*. *The watch projects* a bright holographic sphere that rotates, displaying the words 'NEW JOB.' …



A healthcare worker, dressed in a white lab coat, skillfully *administers a vaccine to a young girl* wearing a patterned top. …



The video features a woman engaged in a makeup tutorial, meticulously *applying a product to her face with a makeup brush*. …

Figure 12. Qualitative comparison on image-to-video task.

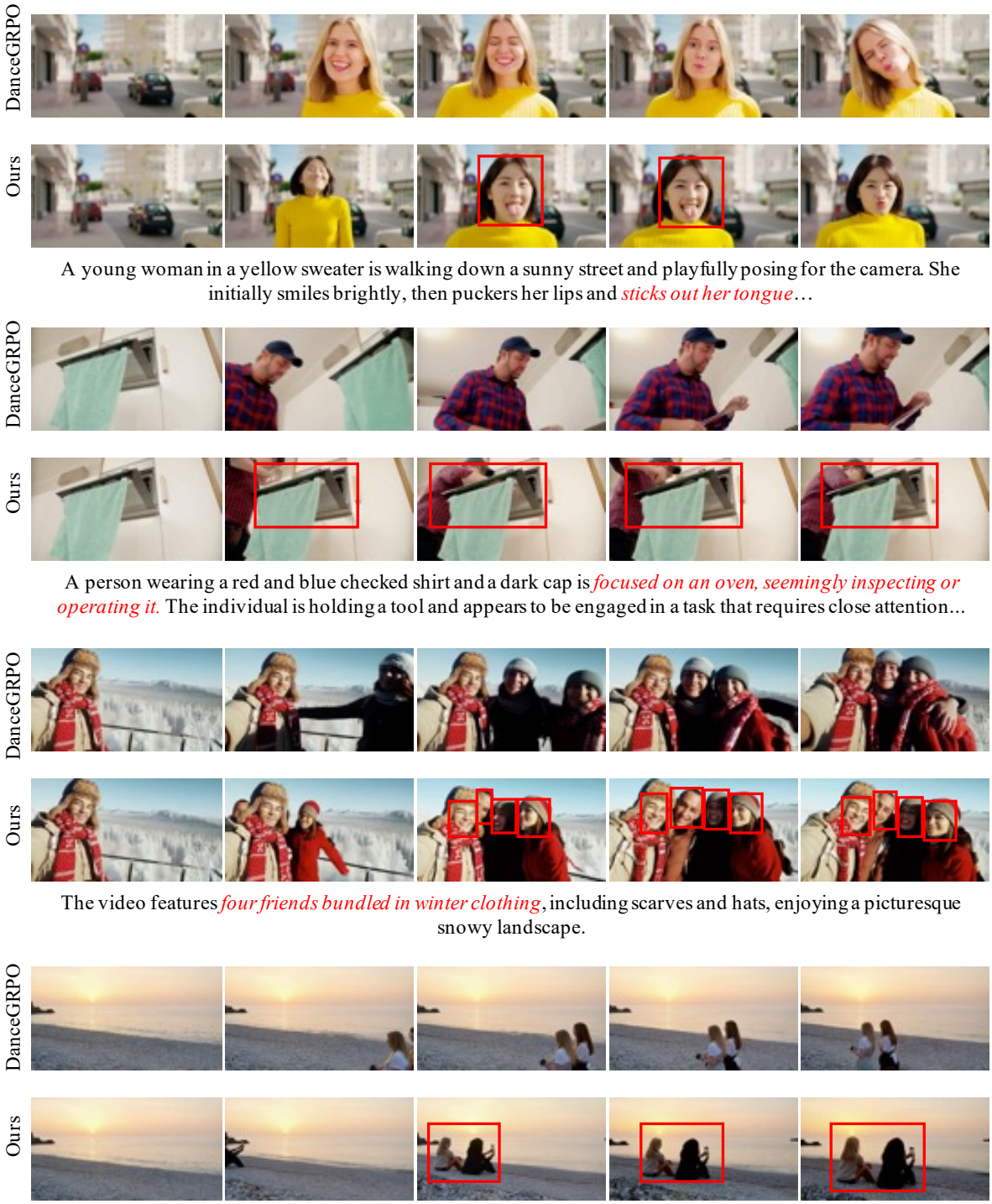


Figure 13. Qualitative comparison on image-to-video task.

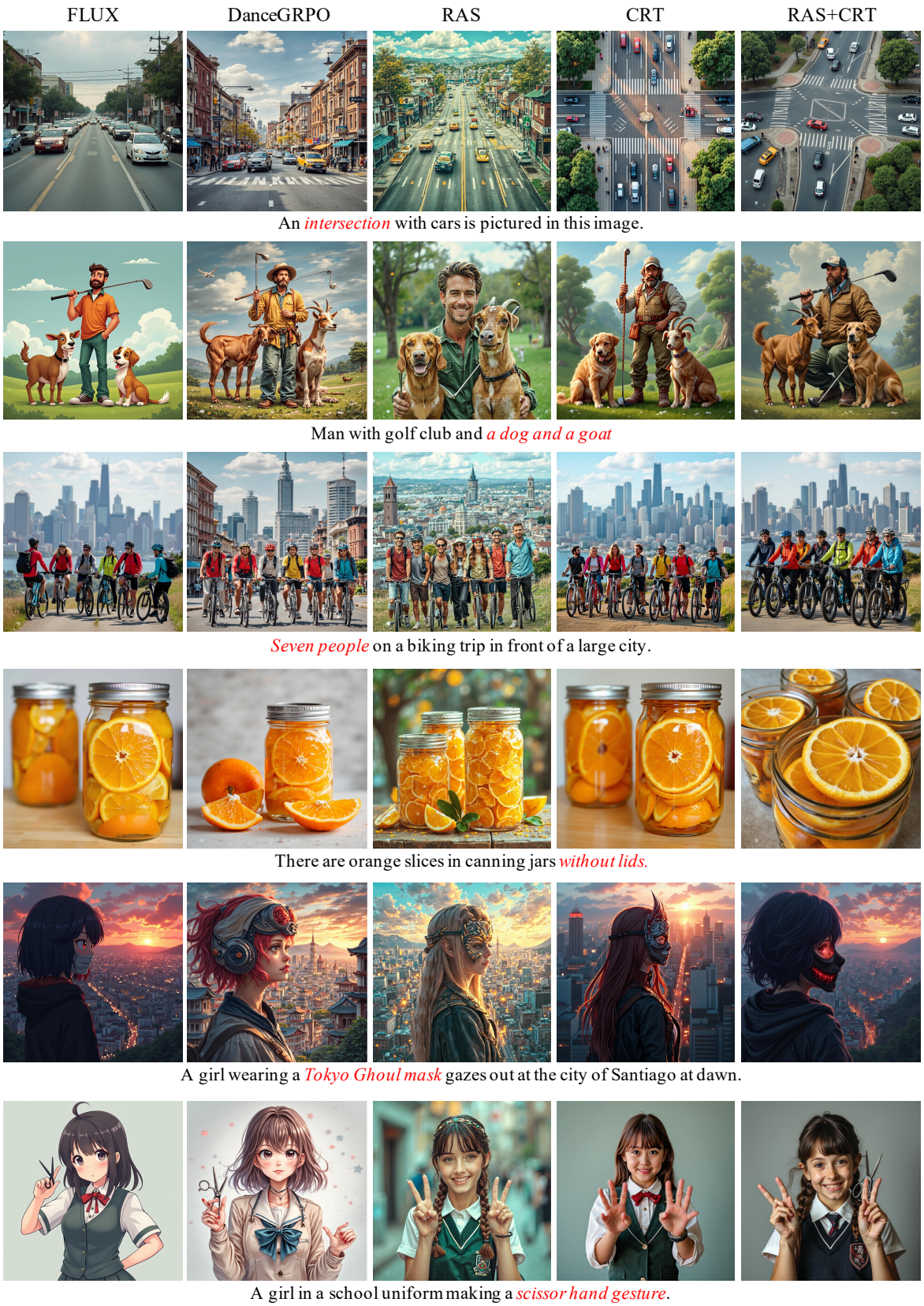


Figure 14. Qualitative comparison on text-to-image task.