DreamRelation: Relation-Centric Video Customization

Supplementary Material

A. Experimental Setup

Datasets. We select 26 types of human interaction videos from the NTU RGB+D Action Recognition Dataset [47, 69] for training. The names of these interactions and their annotated textual descriptions are provided in Tab. 11.

Baselines. Due to the current lack of relational video customization methods, we consider four baselines and detail the implementation of each method below: *1)* Base Model Mochi [77]. We input the test text prompts into the original Mochi for inference and evaluate the results. *2)* Direct LoRA Fine-tuning. We insert LoRAs into all the Query, Key, Value matrices, and FFNs in Mochi for training and inference. The training iterations are set to 1,000. Other training settings, such as the optimizer and LoRA rank, are the same as those in our DreamRelation. *3)* ReVersion [36]. As ReVersion is designed for relational image customization and cannot be directly applied for video generation, we adapt ReVersion to the base model Mochi based on their official code[1]. The training settings follow the default settings provided in the official ReVersion paper. *4)* MotionInversion [83]. Given that MotionInversion is designed based on the Temporal Attention layers within the UNet architecture, and such layers are absent in the MM-DiT architecture, we adapt MotionInversion to Mochi using their official code[2]. Specifically, we integrate the two embeddings from MotionInversion into the query, key, and value matrices of full attention, adhering to their official paper. The learning rate is set to 2e-4, and the weight decay is set to 0.01. The training iterations are 3,000, with other settings consistent with our method. During inference, we utilize the differencing operation from their official paper to mitigate the appearance biases in motion embeddings.

Evaluation metrics. We detail the proposed Relation Accuracy metric utilizing Vision-Language Models (VLMs). Specifically, we input all generated videos into Qwen-VL-Max [2], the state-of-the-art Visual Question Answering (VQA) model, to determine if the generated video conforms to the specified relation, prompting it to return either "yes" or "no." Directly inputting an entire 61-frame video into the VLM would require significant resources and slow response times. To address this, we evenly extract five key frames from each video, including the first and last two frames, and input them into the VLM. The text input template for the VLM is: "Based on the keyframes of the video, analyze whether the two subjects are performing human-like {} interactions. The answer should be 'yes' or 'no'." The "{}" is replaced with a specific relation name, such as "handshaking", for evaluation. We test all videos ten times, count the responses for all videos, convert these into percentages of relation accuracy, and compute the average accuracy as the Relation Accuracy score.

[1] https://github.com/ziqihuangg/ReVersion

[2] https://github.com/EnVision-Research/MotionInversion

B. More Results

Details about the user study. We conduct a user study involving 180 groups of videos with 15 randomly selected relations. Participants are presented with three sets of questions for each of the four anonymous methods, paired with a reference video and a textual prompt. For each group of four generated videos, participants are asked the following questions: (1) Relation Alignment: "Which interaction exhibited in videos is more consistent with the reference video?"; (2) Text Alignment: "Which video better matches the text description?"; and (3) Overall Quality: "Which video exhibits better quality and minimal flicker?". The results of the user study are illustrated in Fig. 7(b).

More qualitative results. To further demonstrate the effectiveness of our DreamRelation, we present additional visual results in Figs. 9 and 10. These examples illustrate the capability of our method to generate videos that align with the specified relations and textual descriptions.

Results of more complex relations. To further validate the generalization capability of our method, we conduct experiments with more complex interactions like riding and fighting, as well as relations among more subjects (>2), as shown in Fig. 11. These results confirm the effectiveness of our approach in handling more complex relations.

Results of HunyuanVideo. The comparison results with HunyuanVideo [42] are presented in Fig. 12 and Table 5. While HunyuanVideo demonstrates improved metrics compared to Mochi, it still struggles with accurately customizing relations, as depicted in Fig. 12. In contrast, our architecture-agnostic method can be seamlessly integrated into the HunyuanVideo model, achieving notable improvements and highlighting its potential to adapt to more advanced models.

Table 5. **Quantitative comparison results of HunyuanVideo.**

Method	Relation Accuracy	CLIP-T	Temporal Consistency
HunyuanVideo	0.3657±0.02	0.3253	0.9905
Ours (HunyuanVideo)	**0.4929**±0.01	**0.3256**	**0.9951**

Results of Training on All 26 Actions. When trained on all 26 relations, our method can still effectively customize

these relations, as shown in Fig. 13, indicating potential for generalization to a tuning-free paradigm.

Comparison with "Image Customization + Image-to-Video" pipeline. Compared to the "Image Customization + Image-to-Video" approach, our method is explicitly designed for video customization, offering better results, end-to-end flexibility, and scalability. Meanwhile, the T2I+I2V paradigm suffers from cumulative errors due to its two-stage nature, limited controllability from image models, and poor relational accuracy. As a result, it fails to generate fine-grained relations like fighting moves, as shown in Fig. 14, even when using a SOTA I2V model. The results in Tab. 6 further demonstrate that our method achieves better performance than this paradigm.

Table 6. **Quantitative comparison results of "Image Customization + Image-to-Video" pipeline.**

Method	Relation Accuracy	CLIP-T	Temporal Consistency
RelationBooth [71] + Wan2.1 I2V [81]	0.3876±0.03	0.3236	0.9947
Ours	**0.4452**±0.01	**0.3248**	**0.9954**

Comparison of training cost. Tab. 7 shows that on Mochi (10B parameters), our approach uses less VRAM and fewer parameters compared to MotionInversion [83], while maintaining comparable training time with 20% fewer steps, yet achieving superior performance. Additionally, our method allows for flexible adjustment of the selection probability of Relation LoRAs to increase their update frequency, thereby reducing training time.

Table 7. **Comparison of VRAM usage, training parameters, and time required using Mochi (10B parameters) on one A100.**

	MotionInversion	Ours	Ours (3 subjects)
VRAM (F=37)	63.0G	39.9G	42.3G
VRAM (F=61)	**OOM**	51.9G	53.2G
Parameters	1566.14M	39.30M	47.18M
Avg Time/step	7.2s	7.5s	7.5s

C. More Ablation Studies

Effects of Loss Lam λ_1. To determine the optimal value for the loss weight λ_1, we vary its value and measure its impact. As shown in Tab. 8, increasing the loss weight of space-time relational contrastive loss results in degradation of Relation Accuracy. We argue that over-emphasizing contrastive learning may ignore detailed information from training videos, leading to degraded performance. Therefore, we set λ_1 to 0.01 for the best performance.

Effects of Mask Lam. To identify the optimal mask weight λ_m, we explore various values and assess their impact. As shown in Tab. 9, both excessively high and low mask weights can result in poor performance. We argue that low mask weights fail to direct the model's focus on the area of interest, while high weights lead to excessive emphasis, causing the neglect of other visual cues. Based on the results, we set λ_m to 50.

Table 8. **Ablation study of the loss weight λ_1.**

λ_1	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
0.01	**0.4452**±0.01	0.3248	**0.9954**	2079.87
0.10	0.3964±0.03	0.3241	**0.9954**	2088.71
1.00	0.2998±0.01	**0.3254**	**0.9954**	**1971.29**

Table 9. **Ablation study of the mask weight λ_m.**

λ_m	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
1	0.3469±0.07	0.2826	0.9942	2294.98
25	0.3899±0.04	0.3185	0.9953	2117.49
50	**0.4452**±0.01	**0.3248**	**0.9954**	2079.87
100	0.4018±0.04	0.3246	0.9952	**2050.10**

Effects of positive and negative numbers. We conduct ablation studies to investigate the effects of varying the number of positive and negative samples in space-time relational contrastive loss. A higher number of positive samples emphasizes the alignment of relational information during training, while an increased number of negative samples focuses more on distinguishing appearance information. We observe that different combinations have varying effects, and based on the experimental results, we chose to set n_{pos} to 4 and n_{neg} to 10.

Table 10. **Ablation study of the number of positive and negative samples.**

n_{pos}	n_{neg}	Relation Accuracy	CLIP-T	Temporal Consistency	FVD↓
1	10	0.3151±0.04	**0.3259**	0.9954	2089.79
1	30	0.2817±0.03	0.3125	**0.9957**	2067.35
1	60	0.3338±0.06	0.3154	0.9954	2113.27
2	10	0.3321±0.02	0.3227	0.9950	2254.62
4	10	**0.4452**±0.01	0.3248	0.9954	2079.87
2	30	0.4378±0.02	0.3168	0.9953	**2009.92**
4	60	0.3793±0.03	0.3237	0.9952	2156.28

D. Limitations

Existing metrics for relation accuracy may not fully capture the customization capabilities of models. While using VLMs simplifies evaluation and reduces bias, the metric relies on VLM's capabilities; future work should develop metrics that align better with human perception.

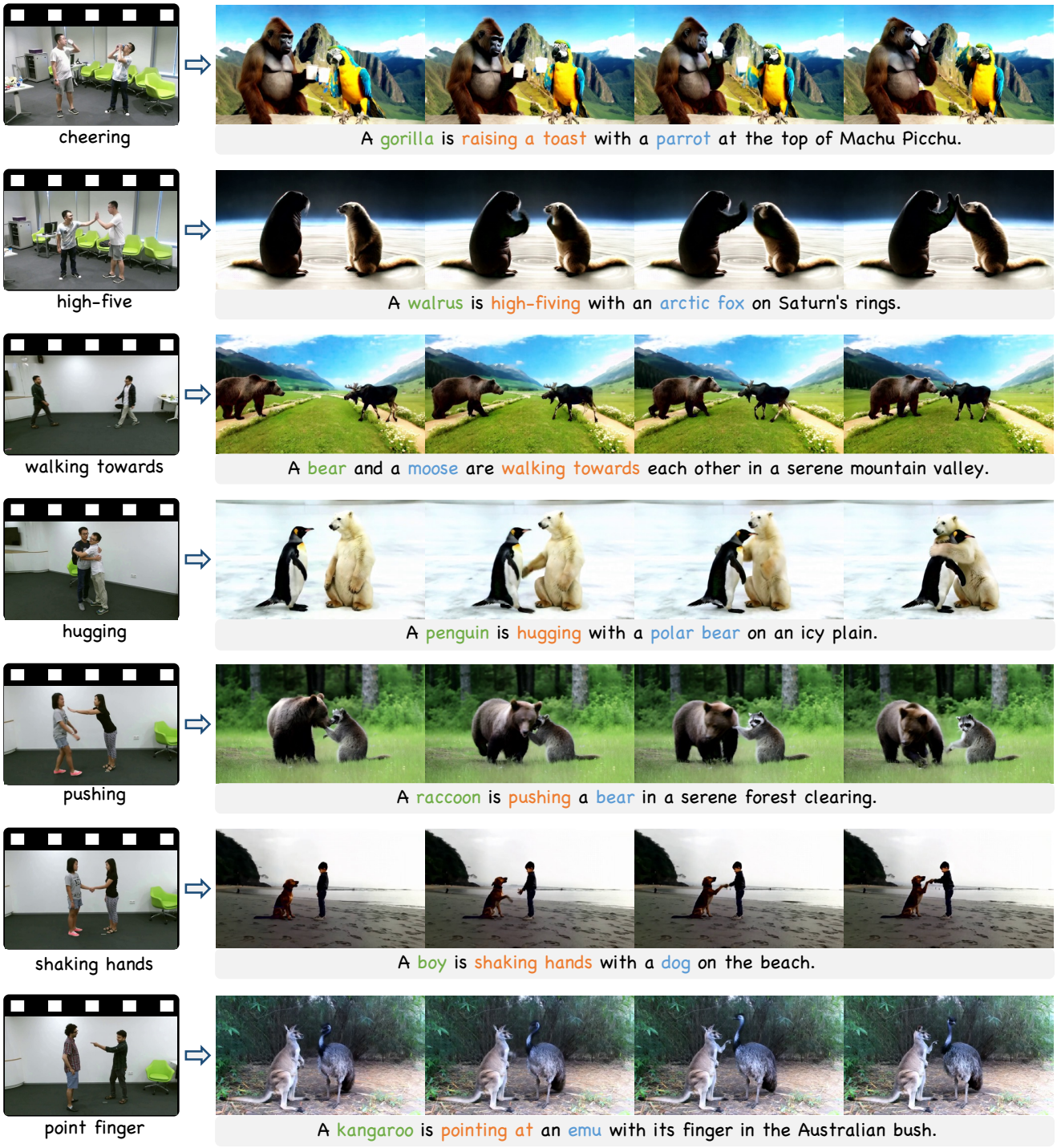

Figure 9. **More qualitative results of DreamRelation (1/2)**. Please zoom in for a better view.

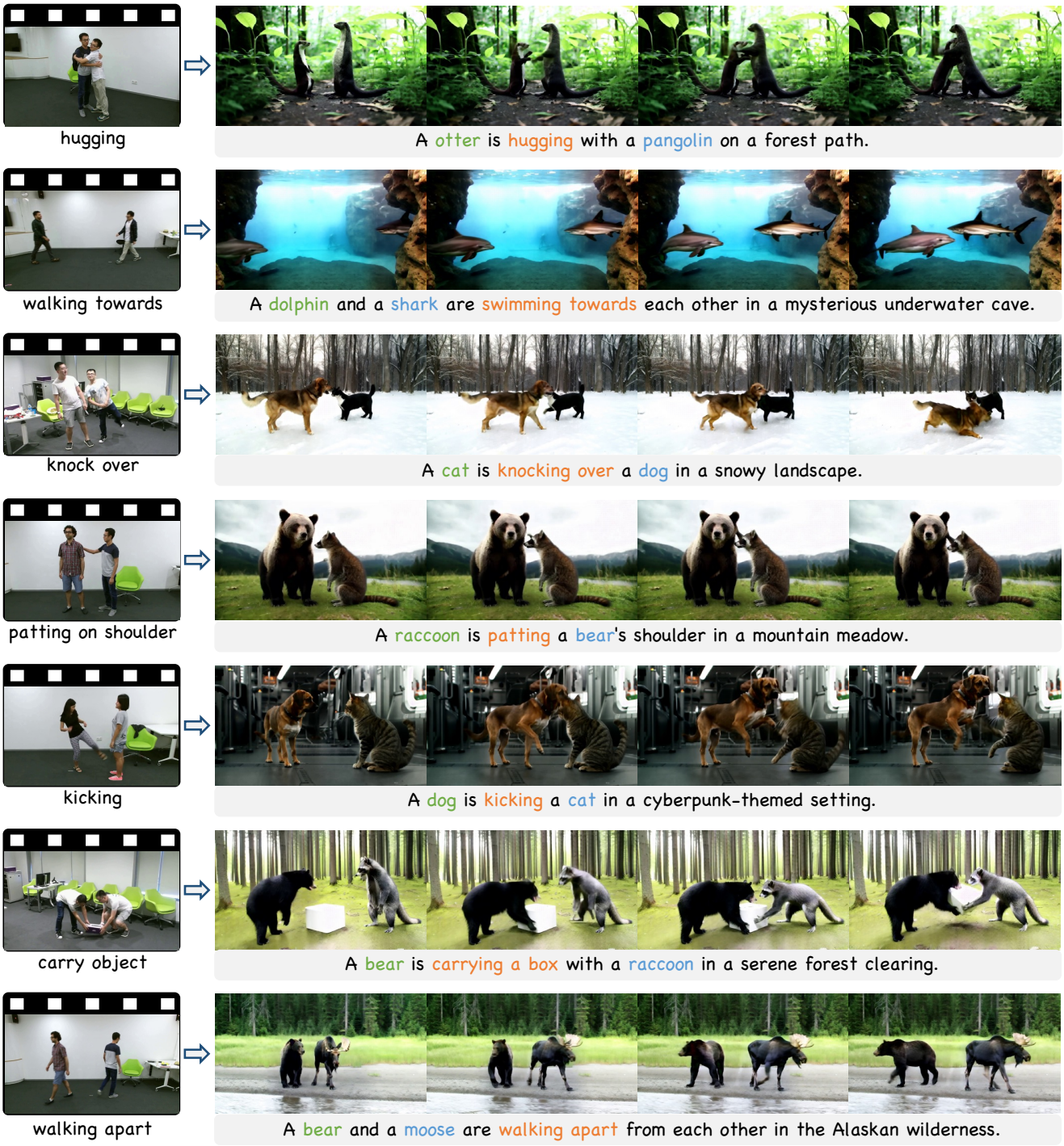

Figure 10. **More qualitative results of DreamRelation (2/2)**. Please zoom in for a better view.

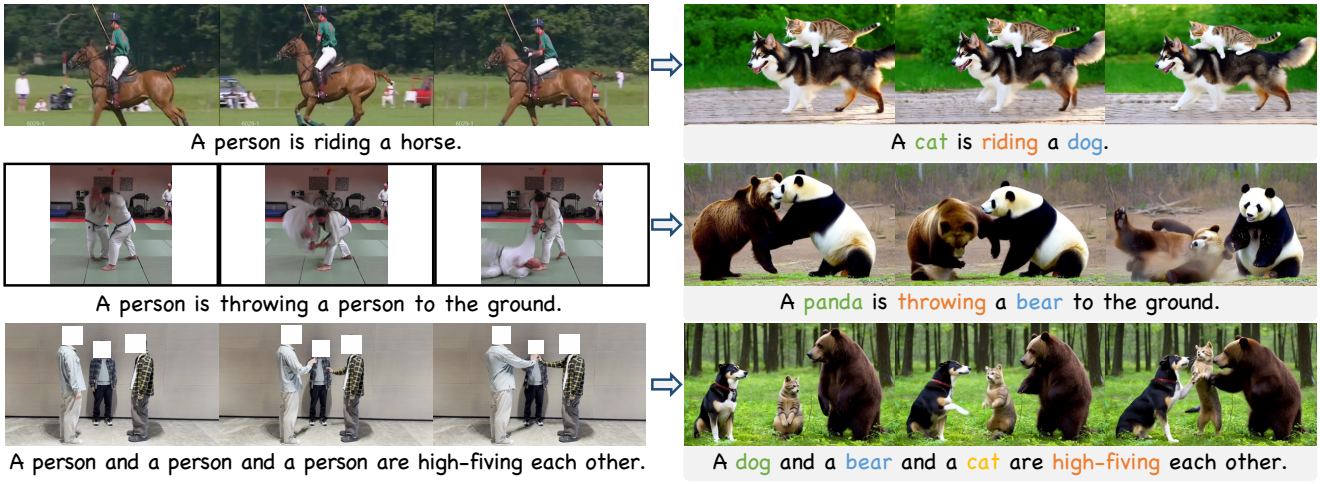

Figure 11. **Qualitative results of DreamRelation on complex relations**. Please zoom in for a better view.

Table 11. The list of 26 human interactions with their textual prompts.

1.	**walking apart**: "A person and a person are walking apart from each other."
2.	**walking towards**: "A person and a person are walking towards each other."
3.	**shaking hands**: "A person is shaking hands with a person."
4.	**hugging**: "A person is hugging with a person."
5.	**point finger**: "A person is pointing at a person with his finger."
6.	**pat on back**: "A person is patting a person's shoulder."
7.	**pushing**: "A person is pushing a person."
8.	**kicking**: "A person is kicking a person."
9.	**punch or slap**: "A person is punching a person."
10.	**rock-paper-scissors**: "A person is playing rock-paper-scissors with a person."
11.	**support somebody**: "A person is supporting a person while walking."
12.	**whisper**: "A person is whispering to a person."
13.	**follow**: "A person is following a person."
14.	**take a photo**: "A person is taking a photo of a person."
15.	**carry object**: "A person is carrying a box with a person."
16.	**cheers and drink**: "A person is raising a toast with a person."
17.	**high-five**: "A person is high-fiving with a person."
18.	**step on foot**: "A person is stepping on a person's foot."
19.	**shoot with gun**: "A person is shooting a person with a water gun."
20.	**knock over**: "A person is knocking over a person."
21.	**giving object**: "A person is giving an object to a person."
22.	**touch pocket**: "A person is touching a person's pocket."
23.	**hit with object**: "A person is hitting a person with an object."
24.	**wield knife**: "A person is wielding a toy knife towards a person."
25.	**grab stuff**: "A person is grabbing an item from a person."
26.	**exchange things**: "A person and a person are exchanging items with each other."

Figure 12. **Qualitative comparison results using HunyuanVideo [42].** Our method is architecture-agnostic and can seamlessly integrate into HunyuanVideo.

Figure 13. **Qualitative results of training on all 26 actions**, indicating our method's potential for generalization to a tuning-free paradigm.

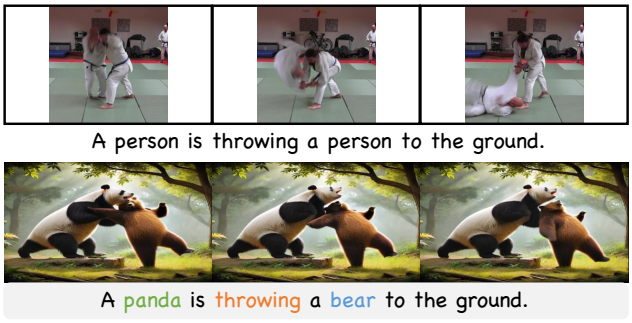

Figure 14. **Qualitative results of "Image Customization + Image-to-Video" pipeline** (RelationBooth [71] + Wan2.1 I2V [81]). This pipeline cannot accurately customize relational patterns.