DualReal: Adaptive Joint Training for Lossless Identity-Motion Fusion in Video Customization

Supplementary Material

1. Experimental Details

This section describes the implementation of our primary experiments and ablation studies. For each method, we provide detailed information on the setup. We list hyperparameter values, data pre-processing and post-processing steps, training schedules, and evaluation protocols. All information is provided to ensure reproducibility and clarity.

DualReal. We run 1,000 training steps for every test case. We set $\gamma = 0.5$ so that each step has a 50% chance of motion training. The learning rate is 1e-3. We use the AdamW optimizer to ensure stable convergence and effective weight regularization. Under these settings, our method consistently produces high-quality customized videos. Each output contains 49 frames at a resolution of 480×720 pixels.

Baseline. For MotionBooth, we adopt LaVie-base as the text-to-video backbone, set the learning rate to 5e-6, train for 300 steps with a batch size of 10 using the unique token "sks" and the AdamW optimizer. For both LoRA and full-parameter fine-tuning, we follow the official CogVideoX training code: LoRA uses a learning rate of 1e-3 with 300 identity steps and 300 motion steps, while full fine-tuning uses a learning rate of 1e-4 with 200 identity steps and 130 motion steps. For DreamVideo, we build on the ModelScopeT2V V1.5 base model and follow the recommended schedule, first training the identity stage for 3000 steps (batch size 4, learning rate 1e-4), then continuing identity training for 500 steps (batch size 4, learning rate 1e-5), and finally running multi-video motion training for 600 steps (batch size 2, learning rate 1e-5).

Prompts. Given a target identity and motion, we employ a large language model to enrich the prompt by appending details, such as clothing styles, accessories, and situating the subject in diverse settings that align with the intended action. This automated prompt expansion introduces both semantic variety and environmental complexity, enabling us to rigorously evaluate the extent to which our customized video framework can accurately interpret and render nuanced textual edits.

2. More Main Results

To highlight the differences among methods, we conduct a comprehensive qualitative comparison between *DualReal* and several state-of-the-art baselines. Whereas prior approaches often sacrifice either identity fidelity or motion realism, *DualReal* delivers both: it preserves distinctive identity features while producing smooth, temporally consistent motion. This dual achievement stems directly from our joint training scheme, which aligns identity and motion objectives within a unified optimization process and thereby resolves the inherent conflicts between static appearance and dynamic behavior. The result is a harmonious fusion of identity and motion, as shown in Fig. 3.

3. More Ablation Results

We provide additional qualitative results in Fig. 4 that further validate the influence of each component, aligning with the descriptions provided in the main paper. Omitting Dual-aware Adaptation introduces visible artifacts, especially around the hands and chin, that markedly degrade clarity. Replacing our StageBlender Controller with direct fusion (i.e., using fixed adapter weights at inference) causes the model to over-adapt to motion dynamics. Eliminating weight grouping so that all blocks receive uniform modulation leads to weakened identity preservation and a loss of background detail. Together, these findings demonstrate that every module in our pipeline is critical for achieving high-quality, customized video generation.

4. More Cases

The *DualReal* framework dynamically tailors its dual processing pathways to any combination of user-supplied identity references and motion sequences, irrespective of their complexity tier. By automatically calibrating to input difficulty—from simple to intricate actions—it synthesizes personalized 720x480 resolution videos comprising 49 temporally consistent frames. Crucially, the system rigorously preserves subject identity characteristics while ensuring smooth motion transitions across all generated content. This dual-path adaptability addresses the core challenge of reconciling visual authenticity with kinematic continuity, establishing a generalized solution for user-customized video generation across diverse input scenarios. We further demonstrate the generation effect of our method on different cases and prompts as shown in Fig. 1 and Fig. 2.

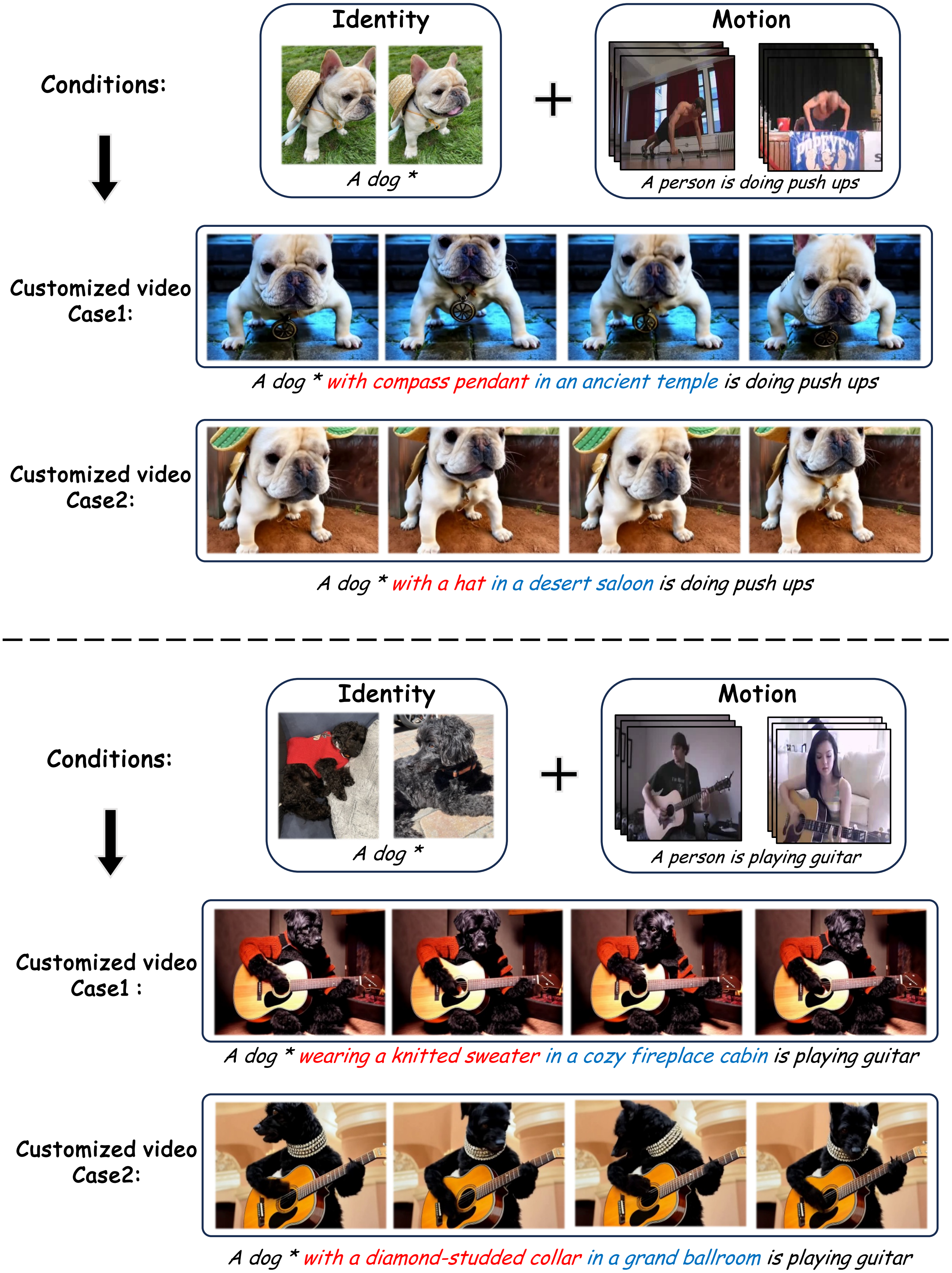
Conditions:
Identity
A dog *
Motion
A person is doing push ups
Customized video Case1:
A dog * with compass pendant in an ancient temple is doing push ups
Customized video Case2:
A dog * with a hat in a desert saloon is doing push ups
Conditions:
Identity
A dog *
Motion
A person is playing guitar
Customized video Case1 :
A dog * wearing a knitted sweater in a cozy fireplace cabin is playing guitar
Customized video Case2:
A dog * with a diamond-studded collar in a grand ballroom is playing guitar

Figure 1. Generated customization results of our proposed novel paradigm ***DualReal***. Given subject images and motion videos, *DualReal* generates high-quality customized identity and motion simultaneously, without compromising the consistency of either dimension.

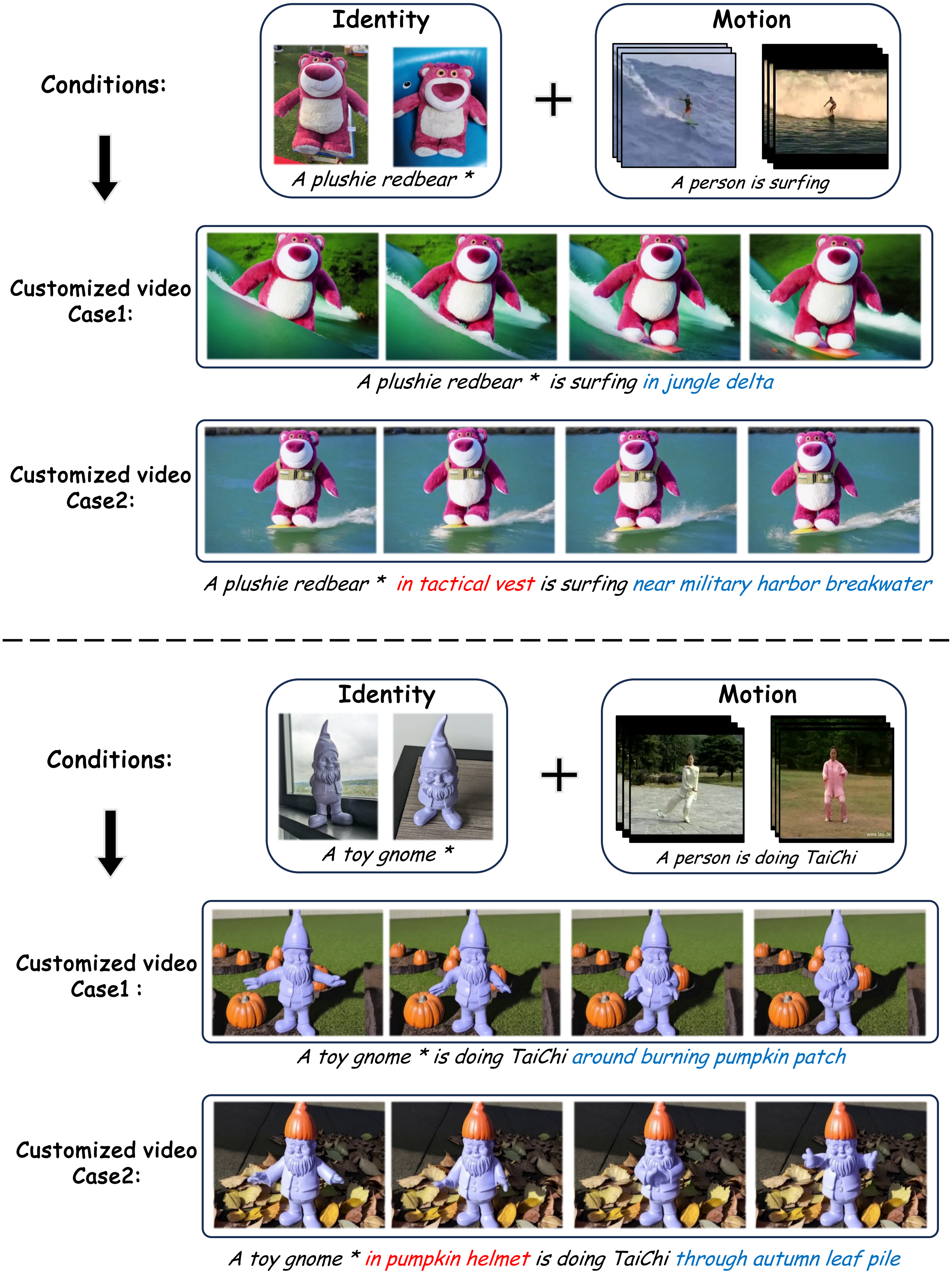
Identity
Motion
Conditions:
A plushie redbear *
+
A person is surfing
Customized video Case1:
A plushie redbear * is surfing in jungle delta
Customized video Case2:
A plushie redbear * in tactical vest is surfing near military harbor breakwater
Identity
Motion
Conditions:
A toy gnome *
+
A person is doing TaiChi
Customized video Case1 :
A toy gnome * is doing TaiChi around burning pumpkin patch
Customized video Case2:
A toy gnome * in pumpkin helmet is doing TaiChi through autumn leaf pile

Figure 2. Generated customization results of our proposed novel paradigm ***DualReal***. Given subject images and motion videos, *DualReal* generates high-quality customized identity and motion simultaneously, without compromising the consistency of either dimension.

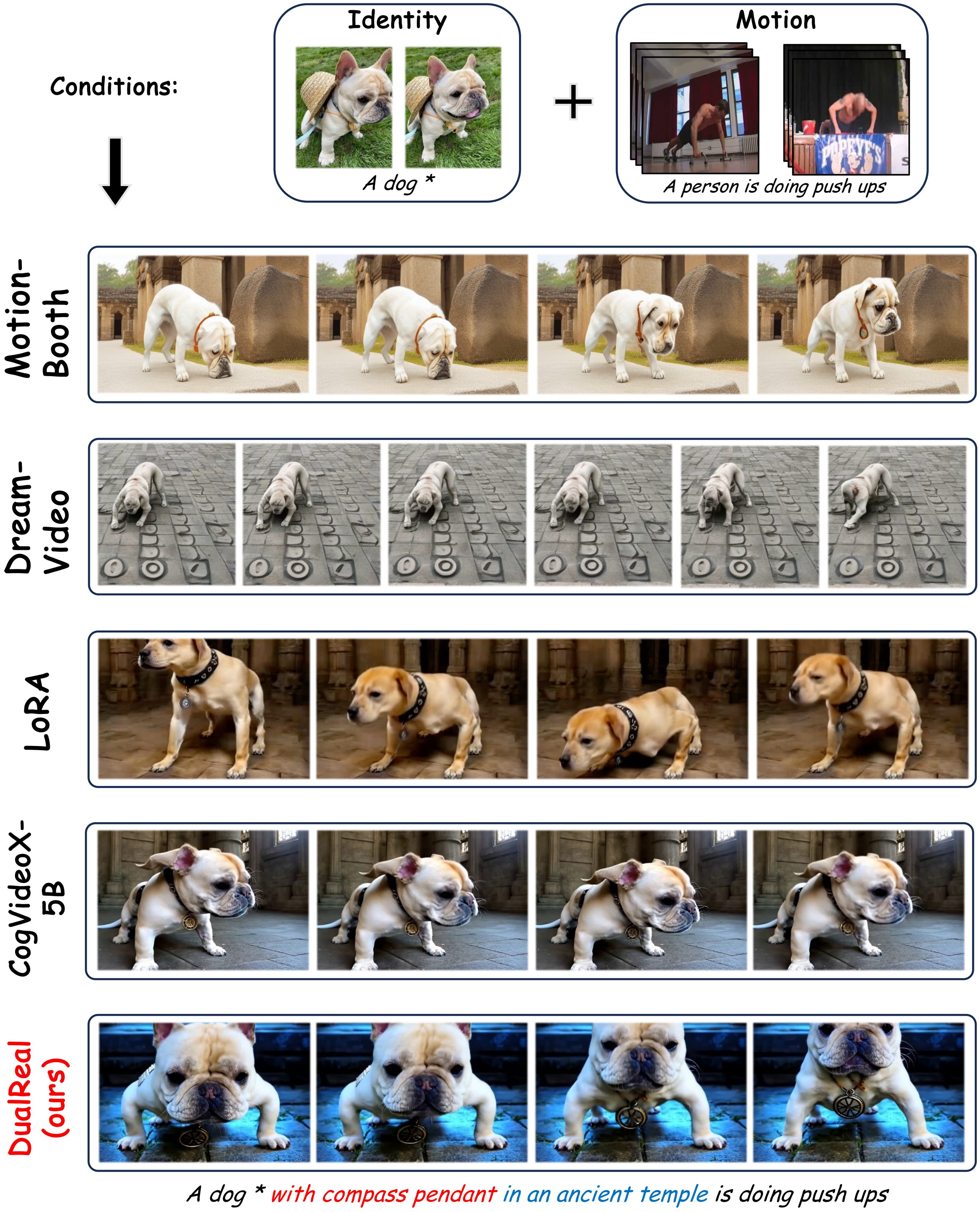
Conditions:
Identity
A dog *
+
Motion
A person is doing push ups
Motion-Booth
Dream-Video
LoRA
CogVideoX-5B
DualReal (ours)
A dog * with compass pendant in an ancient temple is doing push ups

Figure 3. **More Qualitative comparison with existing methods.** The result shows that while MotionBooth maintains identity fidelity, it fails to model motion patterns effectively. DreamVideo suffers from pattern conflicts during inference, resulting in inconsistent identity. Similarly, CogVideoX-5B and LoRA struggle to preserve identity due to their decoupled training methods. In contrast, DualReal achieves high identity consistency with coherent motion, demonstrating the advantage of joint training in balancing pattern conflicts.

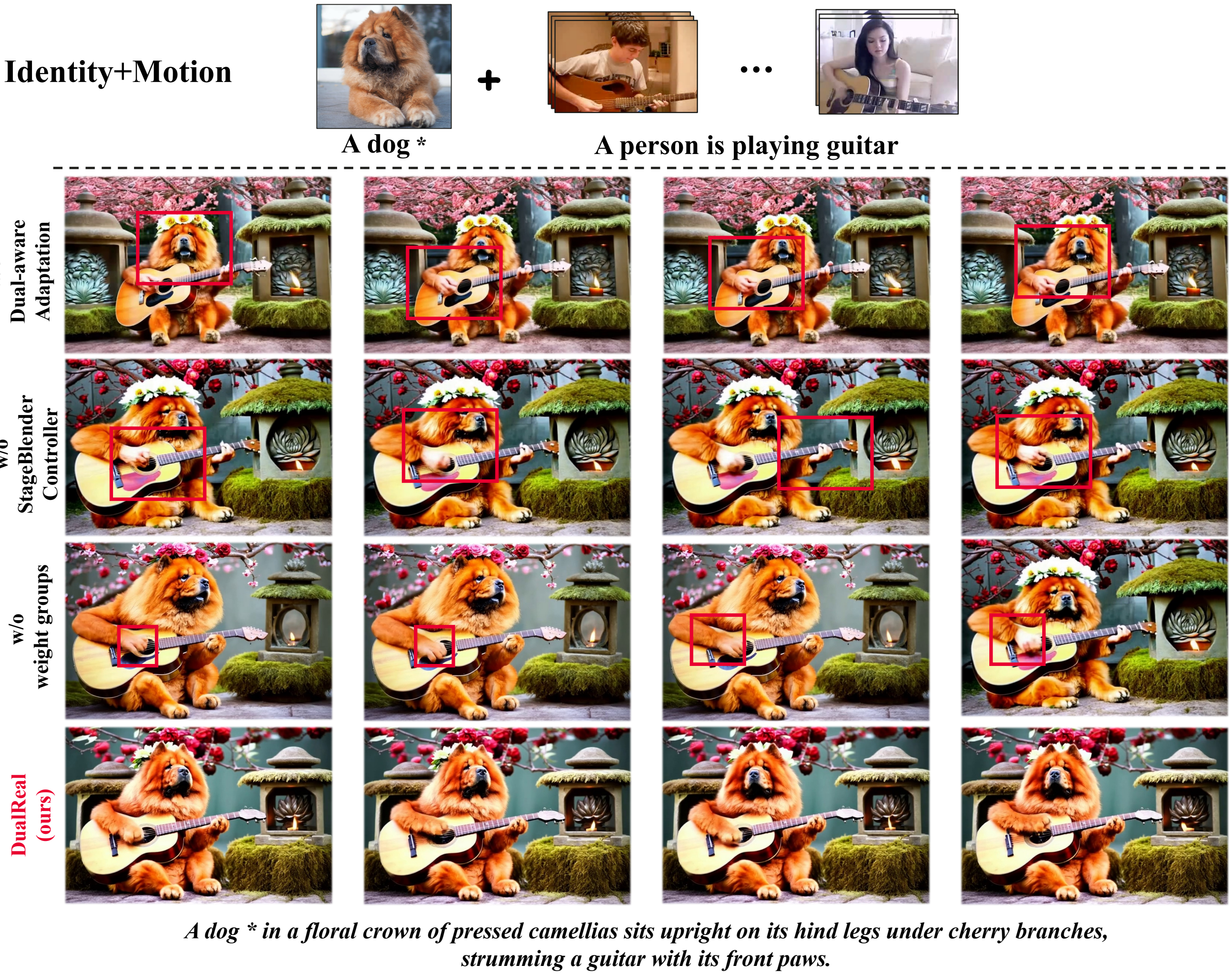
Identity+Motion
A dog *
A person is playing guitar
w/o Dual-aware Adaptation
w/o StageBlender Controller
w/o weight groups
DualReal (ours)
A dog * in a floral crown of pressed camellias sits upright on its hind legs under cherry branches, strumming a guitar with its front paws.

Figure 4. **Qualitative ablations studies on each component.** Omitting Dual-aware Adaptation introduces artifacts on the subject's hands, significantly reducing clarity. Moreover, using fixed weights for the dimensional adapters without the StageBlender Controller causes over-adaptation to the motion pattern, and omitting weight grouping further undermines identity fidelity.