Learning Disentangled Identifiers for Action-Customized Text-to-Image Generation

Supplementary Material

A. Benchmark Details

In this section, we describe the presented **ActionBench** in detail. The full benchmark will be publicly available.

A.1. Actions

We define eight diverse, unique and representative actions as follows:

- **salute**: “*salutes*”
- **gesture**: “*raises one finger*”
- **cheer**: “*raises both arms for cheering*”
- **pray**: “*has hands together in prayer*”
- **sit**: “*sits*”
- **squat**: “*squats*”
- **meditate**: “*meditates*”
- **handstand**: “*performs a handstand*”

where the action categories (displayed in **boldface**) are used only to distinguish between actions, and the actions can be best described with the exemplar images. And the text descriptions (displayed in *italics*) that are used for Stable Diffusion are obtained using an image captioning model.

A.2. Subjects

We provide 23 subjects for evaluation as follows:

- **generic human**: “*A boy*”, “*A girl*”, “*A man*”, “*A woman*”, “*An old man*”
- **well-known personalities**: “*Barack Obama*”, “*Michael Jackson*”, “*David Beckham*”, “*Leonardo DiCaprio*”, “*Messi*”, “*Spiderman*”, “*Batman*”
- **animals**: “*A dog*”, “*A cat*”, “*A lion*”, “*A tiger*”, “*A bear*”, “*A polar bear*”, “*A fox*”, “*A cheetah*”, “*A monkey*”, “*A gorilla*”, “*A panda*”

where diverse and unseen subjects and the introduction of animals demand that, models not only retain pre-trained knowledge without forgetting, but also accurately generate animal representations without distortion or anomalies.

B. Baseline Details

All baselines use the prompt template provided by the ActionBench. Each prompt details its image content, leaving the action blank for filling with identifiers from different methods. Other details are:

- **ControlNet** [35]: We use OpenPose [1] as a preprocessor to estimate the human pose of the given reference image.
- **DreamBooth** [25]: The training is with a batch size of 2 and a learning rate of 5e-5. The number of training steps is set to 1000, and 50 images are generated for prior preservation.

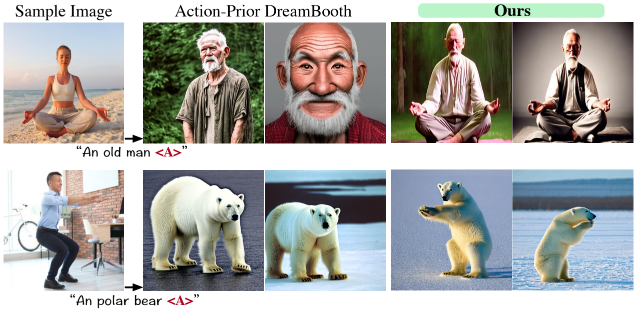

Figure 9. **Comparison with action-prior DreamBooth.** This extended DreamBooth still struggles with inverting action features.

- **Textual Inversion** [5]: The training is with a batch size of 2 and a learning rate of 2.5e-4. The number of training steps is set to 3000.
- **ReVersion** [7]: The training is with a batch size of 2 and a learning rate of 2.5e-4. The number of training steps is set to 3000. The weighting factors of the denoising loss and the steering loss are set to 1.0 and 0.01. The temperature parameter in the steering loss is set to 0.07. And in each iteration, 8 positive samples are randomly selected from the basis preposition set.
- **Custom Diffusion** [9]: The training is with a batch size of 2 and a learning rate 1e-5. The number of training steps is 2000. And the number of regularization images is 200.
- **P+** [30]: The training is with a batch size of 8 and a learning rate 5e-3. The number of training steps is 500.

C. Additional Experimental Results

C.1. Comparison with Action-Prior DreamBooth

Our ADI utilizes the generated action-different samples with the same context to capture the context-related features. To analyze the advantages of controlling updates with these data rather than directly employing them in training, we present a new baseline named action-prior DreamBooth, which replaces the class prior generated by original Stable Diffusion with these action-different samples. Therefore, in addition to the inherent action invariance, contextual invariance also emerges in the training data. However, as shown in Fig. 9, this new baseline still struggles with inverting action-specific features. This observation suggests a lack of ability to capture high-level invariance.

C.2. Generalization Across Diverse Styles

ADI is designed to separate and inverse abstract the action concepts from the details of subjects and objects, background, color, or style in user images. This allows the generation images to generalize to specific styles through prompting, shown as Fig. 10.

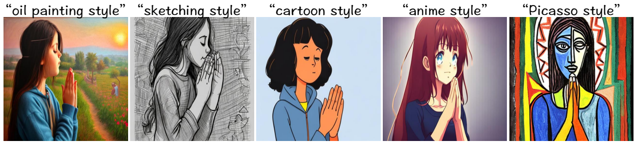

Figure 10. **ADI can generate images with different styles by prompting.** The original prompt is "*A girl* <A>" where "<A>" represents the action **pray**.

C.3. Visualization of Cross-Attention Maps

To explain why certain channels can be interpreted as "action-related", we visualize the cross-attention maps related to the learned identifiers in Fig. 11. As observed, the learned identifiers focus more on the contour information of the actions rather than the human body. This indicates that ADI avoids reversion on appearance information, thereby enabling generalization to different subjects.

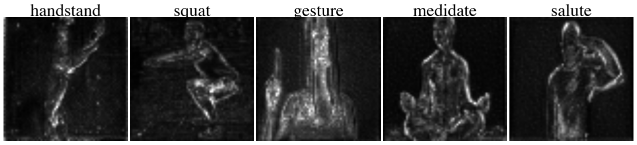

Figure 11. **Visualization of cross-attention maps associated with the learned identifiers.**

C.4. Visualization of Action-Different Pairs

We present the generated images in the action-different pairs in Fig. 12 for reference. Using only a single image for training, the subject-driven model can change actions while preserving contextual information as much as possible. Although the quality of the image may be insufficient, it does not hinder the final inversion of action knowledge.

C.5. Additional Qualitative Results

To show the effectiveness of ADI, we illustrate additional generation results in Fig. 13, covering all actions within ActionBench. The generated images maintain the same action while offering a rich diversity, indicating that the learned identifiers contain solely action information and do not encapsulate irrelevant contextual details such as background, appearance, or even orientation.

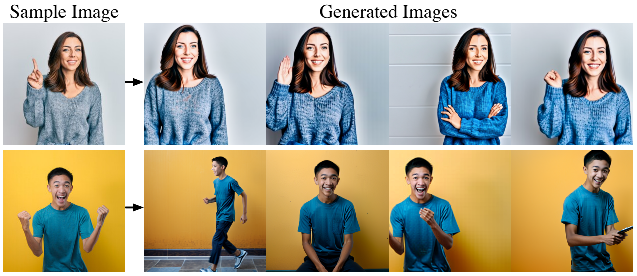

Figure 12. **Visualization of subject-driven generation results for action-different pairs.**

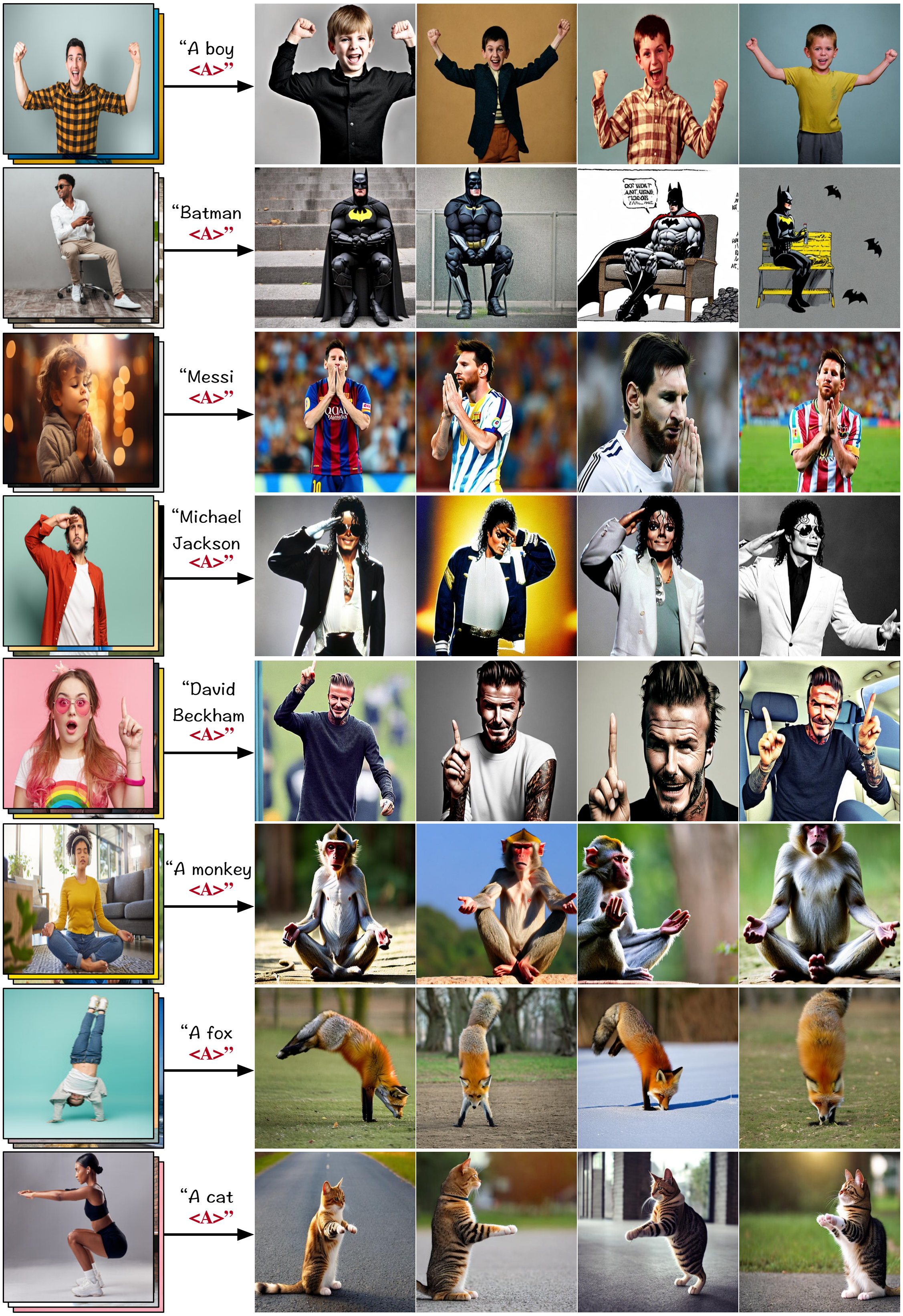

Figure 13. **Additional generation results by ADI, encompassing all the actions within ActionBench.**