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

Siteng Huang Biao Gong Yutong Feng Xi Chen Yuqian Fu Yu Liu Donglin Wang✉

Zhejiang University & Alibaba Group & ETH Zürich & MiLAB, Westlake University

CVPR SEATTLE, WA JUNE 17-21, 2024

Overview

Research Question

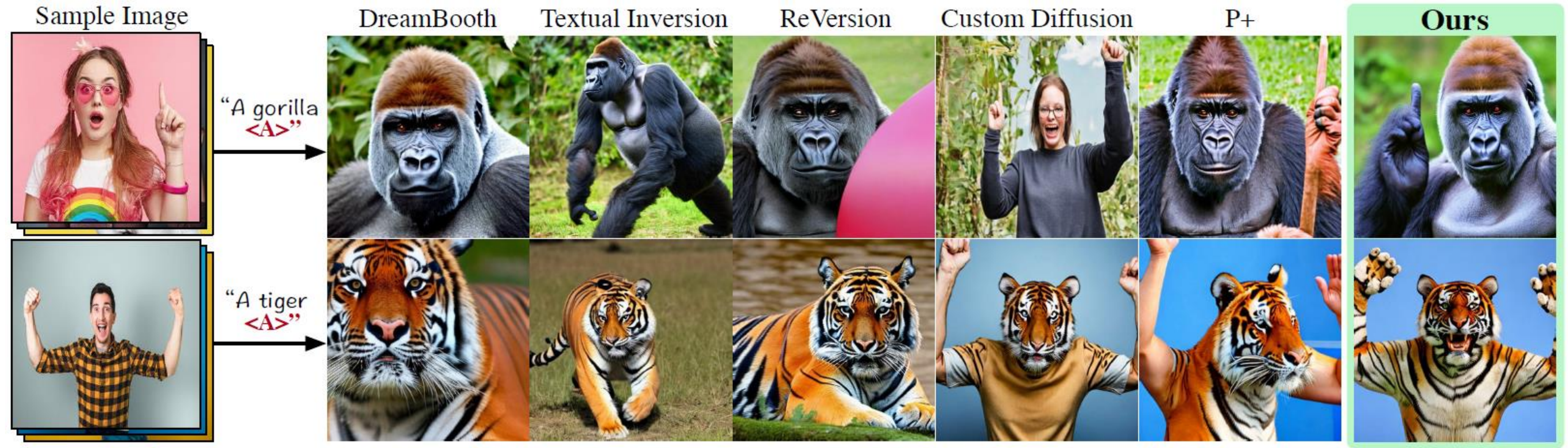
Can we learn the co-existing action from limited data (e.g., 10 images) and generalize it to unseen humans or even animals, without extracting skeleton and sacrificing generation flexibility, diversity, and quality?

Contributions

- propose a novel action customization task, which requires learning the desired action from limited data for future generation.
- contribute the ActionBench, where a variety of unique actions with manually filtered images provide the evaluation conditions for the task.
- devise the Action-Disentangled Identifier (ADI) method, which inverts action-related features into the learned identifiers that can be freely combined with various characters and animals to generate high-quality images.

Background

Action Customization Results by Adapting Subject-Driven Solutions:

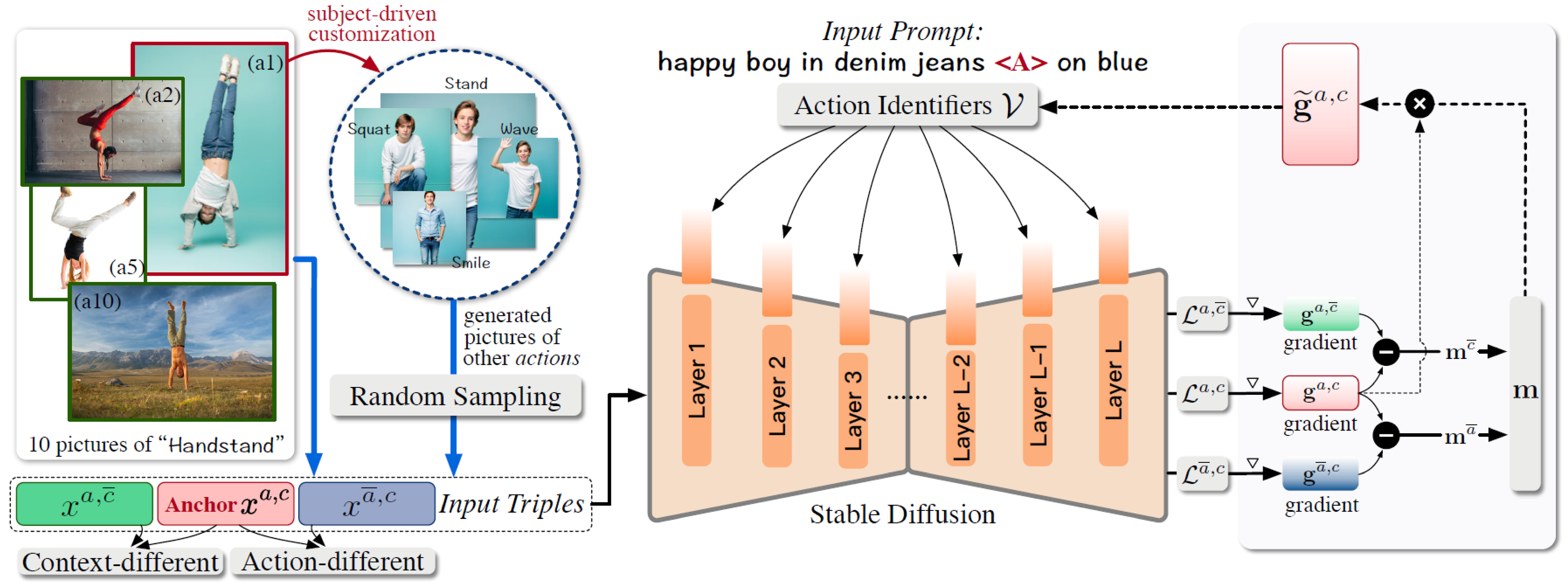


Two Observations:

- **Neglect of high-level action features**: Several methods (DreamBooth, Textual Inversion, and ReVersion) generate images that are unrelated to specific actions, suggesting that they fail to capture the representative characteristics of the actions.
- **Semantic contamination**: Other methods (Custom Diffusion, P+) are capable of encoding action-related knowledge, but they fail to decouple the focus from action-agnostic features, such as the appearance of the human body.

Textual Inversion: $\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(x),\mathbf{y},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{y})\|_2^2\right]$

Action-Disentangled Identifier (ADI)



Expanding Semantic Inversion:

- overcomes the preference to low-level appearance features.
- applies layer-wise identifier tokens to increase the accommodation of various features.

Learning Gradient Mask with Context-Different Pair:

- prevents the identifiers from inverting action-agnostic features in a gradient level.
- given $x^{(a,c)}$ as an *anchor* sample, randomly samples $x^{(a,\bar{c})}$ with the same action but different context (*i.e.*, human appearance and background).
- calculates the absolute value of the difference between the two gradients, where the channels with a small difference can be regarded as action-related and expected to be preserved:

$$\mathbf{g}^{(a,c)} = \frac{\partial\mathcal{L}^{(a,c)}}{\partial\mathbf{v}} \qquad \triangle\mathbf{g}^{\bar{c}} = |\mathbf{g}^{(a,c)} - \mathbf{g}^{(a,\bar{c})}|$$

$$\mathbf{g}^{(a,\bar{c})} = \frac{\partial\mathcal{L}^{(a,\bar{c})}}{\partial\mathbf{v}} \qquad \mathbf{m}_k^{\bar{c}} = \begin{cases} 0, & \triangle\mathbf{g}_k^{\bar{c}} \geqslant \gamma^\beta \\ 1, & \triangle\mathbf{g}_k^{\bar{c}} < \gamma^\beta \end{cases}$$

Learning Gradient Mask with Action-Different Pair:

- uses the single anchor sample to quickly train a subject-driven customization model, and generates $x^{(\bar{a},c)}$ with the same context but different action.
- channels with small gradient difference can be regarded as context-related and expected to be masked:

$$\mathbf{g}^{(\bar{a},c)} = \frac{\partial\mathcal{L}^{(\bar{a},c)}}{\partial\mathbf{v}} \qquad \triangle\mathbf{g}^{\bar{a}} = |\mathbf{g}^{(a,c)} - \mathbf{g}^{(\bar{a},c)}| \qquad \mathbf{m}_k^{\bar{a}} = \begin{cases} 0, & \triangle\mathbf{g}_k^{\bar{a}} < \lambda^\beta \\ 1, & \triangle\mathbf{g}_k^{\bar{a}} \geqslant \lambda^\beta \end{cases}$$

Merging Gradient Masks for Context:

- due to the noise introduced by context variations, identifying action-relevant channels using only context-different or action-different pairs would be difficult and unreliable.
- keeps only the intersection of the unmasked channels as unmasked, and overwrites the gradient of the anchor sample:

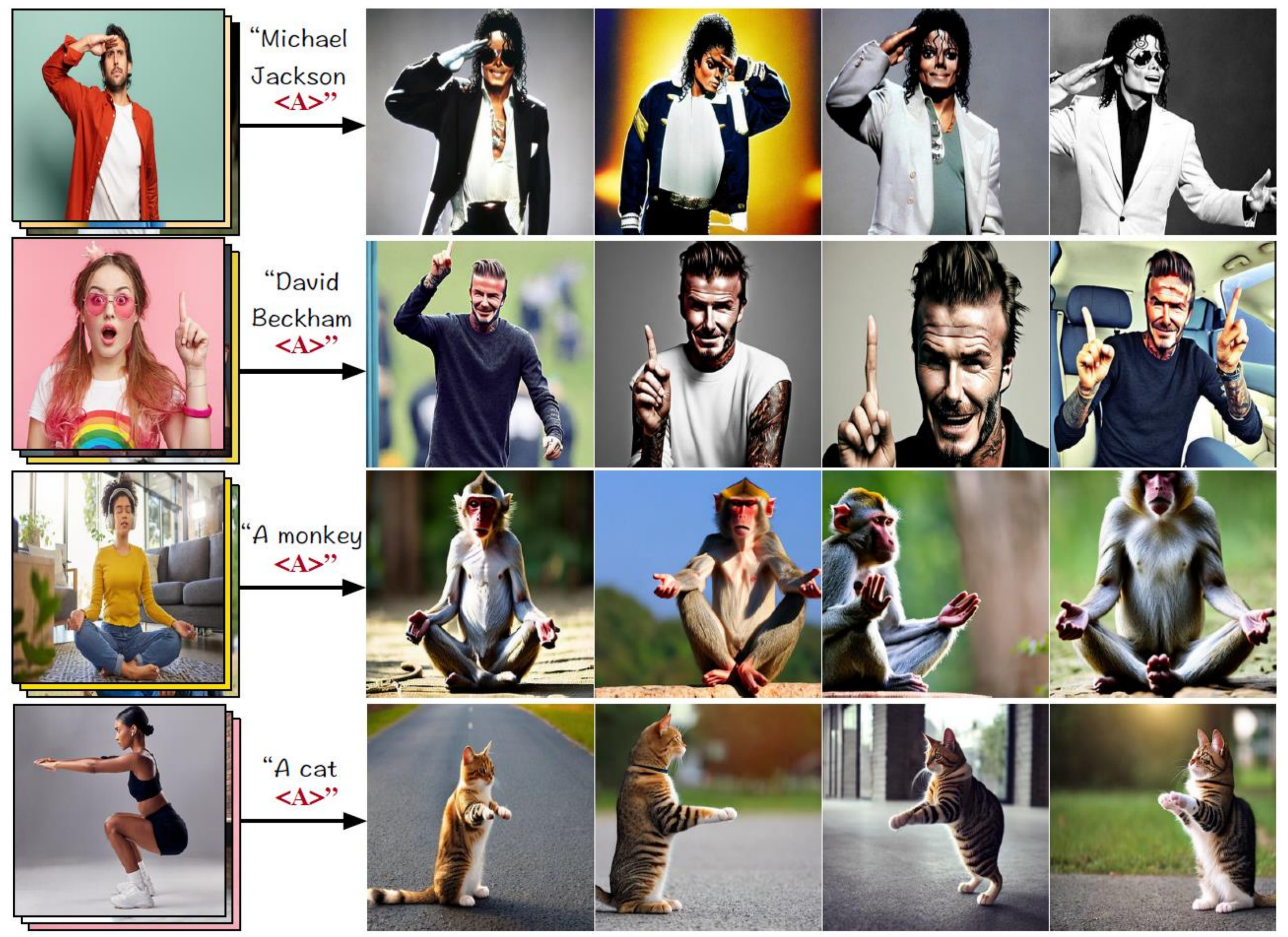
$$\mathbf{m} = \mathbf{m}^{\bar{c}} \cap \mathbf{m}^{\bar{a}} \qquad \tilde{\mathbf{g}}^{(a,c)} = \mathbf{m} \odot \mathbf{g}^{(a,c)}$$

Main Results with Stable Diffusion 2.1

ActionBench:

- 8 unique actions, ranging from single-handed to full-body movements.
- 10 images with textual descriptions for each action.
- 23 subjects for evaluation, including generic humans, well-known personalities, and animals.

Methods	Action	Subject	Total
Stable Diffusion [22]	30.71	84.51	27.17
ControlNet [35]	41.30	42.66	19.29
DreamBooth [25]	2.45	**95.65**	2.45
Textual Inversion [5]	2.17	86.14	1.90
ReVersion [7]	1.63	84.51	1.63
Custom Diffusion [9]	29.62	53.53	7.07
P+ [30]	26.90	80.16	20.92
ADI (Ours)	**60.33**	85.87	**51.09**



For more experimental results, please refer to our paper.

Visit https://adi-t2i.github.io/ADI/ for more information.