Preventing Shortcuts in Adapter Training via Providing the Shortcuts

Anujraaj Argo Goyal **Guocheng Gordon Qian** * **Huseyin Coskun** **Aarush Gupta**

Himmy Tam **Daniil Ostashev** **Ju Hu** **Dhritiman Sagar**

Sergey Tulyakov **Kfir Aberman** **Kuan-Chieh Jackson Wang** *

Snap Inc., https://snap-research.github.io/shortcut-rerouting/

Abstract

Adapter-based training has emerged as a key mechanism for extending the capabilities of powerful foundation image generators, enabling personalized and stylized text-to-image synthesis. These adapters are typically trained to capture a specific target attribute, such as subject identity, using single-image reconstruction objectives. However, because the input image inevitably contains a mixture of visual factors, adapters are prone to entangle the target attribute with incidental ones, such as pose, expression, and lighting. This spurious correlation problem limits generalization and obstructs the model's ability to adhere to the input text prompt. In this work, we uncover a simple yet effective solution: provide the very shortcuts we wish to eliminate during adapter training. In *Shortcut-Rerouted Adapter Training*, confounding factors are routed through auxiliary modules, such as ControlNet or LoRA, eliminating the incentive for the adapter to internalize them. The auxiliary modules are then removed during inference. When applied to tasks like facial and full-body identity injection, our approach improves generation quality, diversity, and prompt adherence. These results point to a general design principle in the era of large models: when seeking disentangled representations, the most effective path may be to establish shortcuts for what should *not* be learned.

1 Introduction

In recent years, text-to-image (T2I) models have undergone remarkable progress, revolutionizing the way we generate and manipulate visual content from natural language prompts [Rombach et al., 2022, Ramesh et al., 2022]. While the expressive power of these foundation models has unlocked myriad creative and practical applications, much of their flexibility is realized not through retraining the backbone itself, but through the introduction of lightweight *adapters*. These adapters—ranging from low-rank adaptation modules (LoRAs) [Hu et al., 2022] to encoders [Ye et al., 2023]—serve as modular steering mechanisms, enabling tailored functionality atop a frozen foundation model. LoRA-based adapters, for instance, have empowered stylized and user-preference-conditioned generation, while encoder-based adapters facilitate personalized synthesis and style injection with impressive specificity [Zhang et al., 2023, Wang et al., 2023, Luo et al., 2024]. In essence, adapter training has emerged as a key enabler of fine-grained control in the modern image generation landscape.

Yet, adapters face a fundamental challenge inherent to their training paradigm. The predominant approach—*single-image reconstruction*, thanks to its simplicity and scalability—asks the adapter to

*Corresponding authors: gqian@snapchat.com, jwang23@snapchat.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 1: **Shortcut Rerouting re-enables text control of pose and expression after adapter training.** In the context of personalized generation, without shortcut rerouting, the adapter overfits to the reference image and reproduces its pose and expression, ignoring the prompt. With Shortcut-Rerouted Adapter Training, the adapter disentangles identity from other factors, allowing the model to respond faithfully to prompt-specified expressions and head poses. This restores compositionality, preserves the prior, and leads to more expressive and diverse generations.

faithfully reproduce a target image from a conditioning signal. An image, as the adage goes, is worth a thousand words; more precisely, an image encodes an entire constellation of attributes—identity, style, geometry, camera parameters, lighting, and beyond. In most cases, however, we wish to learn to encode only some specific attributes, and a thousand words are simply too many. The reconstruction loss, being agnostic to this distinction, indiscriminately incentivizes the adapter to reproduce *all* visual factors present in the image. As a consequence, the adapter entangles the target factor with myriad incidental ones (i.e. *shortcuts*). See Fig. 2. An identity adapter intended to inject only the subject/person's appearance undesirably also copies and pastes their expression, pose, and leaks lighting or background style. Nowhere is this conflation more problematic than in facial personalization, where isolating immutable appearance traits from mutable factors like head pose or expression proves difficult. Moreover, as another key compounding factor, the distribution of the finetuning dataset is often significantly different from that of the foundation model. Such copy-and-paste adapter training often introduces artifacts, such as degraded background generation, distorted human anatomy, and reduced esthetic quality (see Fig. 1). Ideally, an identity encoder must only inject identity.

Our central idea is simple: to prevent the adapter from learning undesirable shortcuts, we explicitly *provide* those shortcuts during training (Fig. 3). Rather than hoping the adapter would disentangle complex factors on its own, we architect the learning process to *route* incidental factors through auxiliary modules—thus relieving the adapter of the burden of accounting for them. This reshapes the optimization landscape: when components of the reconstruction target are already explained by dedicated controllers (e.g., respective modules handling distribution shifts, pose, or expression), the adapter has no incentive to duplicate that behavior. The result is a principled factorization of

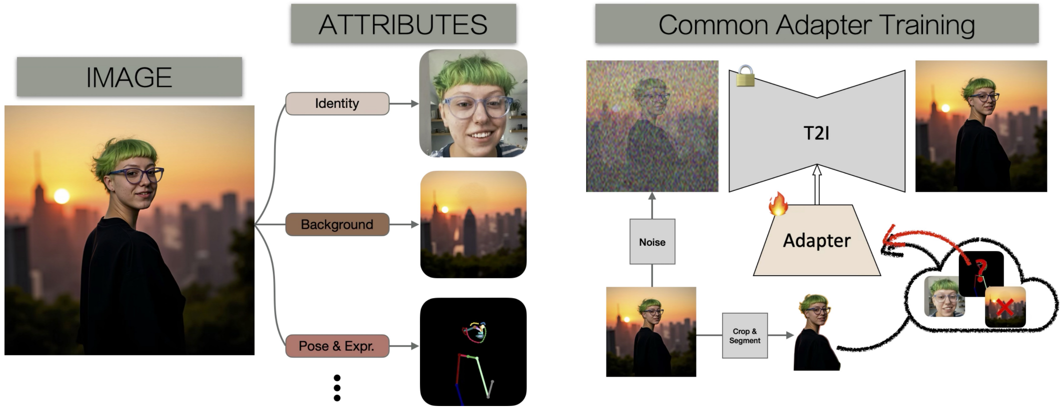

Figure 2: **Common adapter training is susceptible to learning undesired shortcuts.** The common *single-image reconstruction* objective used in adapter training inadvertently encourages the adapter to pick up all the attributes in the adapter input (e.g. pose, expression, background, distribution) and leak them into the generation. While some confounding attributes like background can be factored out using masking, many other cannot. This makes learning a pure "identity" adapter challenging.

responsibility, wherein each module specializes in its designated role. Whereas a naive adapter tends to copy pose and expression directly from the input image—thereby reducing the fidelity of the model prior and leading to degraded background generation—an adapter trained with shortcut rerouting learns to inject only the target identity. This restores prompt-based control over pose and expression and improves prior preservation (Fig. 1). To that end, this approach not only improves compositionality but also enhances the overall realism and diversity of the generated images.

In summary, our **contributions** are as follows:

1. We propose a simple yet effective training paradigm, *Shortcut Rerouting*, for adapter training for large text-to-image (T2I) models.
2. We apply Shortcut Rerouting to the task of personalized image generation, addressing confounding factors such as distribution shift and spurious correlations. We demonstrate two practical instantiations of this idea using well-established tools—LoRA and ControlNet—to explicitly factor out these shortcuts.
3. We empirically validate Shortcut-Rerouted adapters in two distinct settings—facial and full-body personalization—and show improved controllability (via text prompts) with respect to expression, head pose, and body pose, as well as stronger prior preservation. This leads to higher overall image quality, fidelity, and naturalness compared to several strong baselines.

2 Related Work

Adapters in T2I Generation & Personalized Generation. The advent of text-to-image (T2I) diffusion models [Ho et al., 2020, Rombach et al., 2022] has spurred a growing interest in modular methods for task-specific and personalized generation [Ruiz et al., 2023, Gal et al., 2022, Voynov et al., 2023]. Central to this movement is the concept of *adapters*—lightweight modules that steer the behavior of a frozen generative backbone. Among these, LoRA [Hu et al., 2022] has emerged as a widely adopted technique, enabling fine-tuning via low-rank parameter updates. For improved inference efficiency, encoder-based adapters, which inject conditioning signals through learned embeddings or attention modulation, have likewise been instrumental in enabling fine-grained control over appearance, style, and compositionality [Ye et al., 2023, Wang et al., 2024, Xiao et al., 2023, Qian et al., 2025a,b]. Yet, a well-known issue of adapter fine-tuning is that it entangles identity injection with other undesired attributes like style, lighting, pose, and expression. In §4, we apply Shortcut Rerouting to a simple adapter, IP-Adapter, and compare it to strong recent baselines including InfU [Jiang et al., 2025], PulID [Guo et al., 2024], and a community implementation of IP-Adapter.

Shortcuts & Spurious Correlation. The phenomenon of *shortcut learning*—where models exploit spurious or unintended correlations in the data to optimize the training objective—has been extensively studied in the broader machine learning literature [Geirhos et al., 2020, Luo et al., 2021]. In vision tasks, shortcuts often manifest when models rely on superficial cues, such as texture or background, instead of learning the intended high-level semantics [Geirhos et al., 2019]. Within the generative modeling community, recent works have noted analogous behaviors: generative models and their adapters may entangle target factors with irrelevant or transient features present in training data, leading to poor generalization and a lack of modularity. Several approaches have been proposed to combat shortcut learning, including data augmentation [Geirhos et al., 2020], causal regularization [Arjovsky et al., 2019], and architectural interventions [Islam et al., 2020]. In the context of T2I generation, methods such as ControlNet [Zhang et al., 2023] explicitly inject structural conditioning (e.g., pose, layout) to guide synthesis, offering a promising avenue for disentanglement. However, to our knowledge, no prior work has systematically leveraged such auxiliary modules during *adapter training* to proactively absorb spurious factors. Our method bridges this gap by architecting a modular training process that reroutes undesired correlations through dedicated controllers, thereby preventing their entanglement in the adapter's representation.

Lastly, the notion of employing stage-wise training—where an auxiliary "training LoRA" is used solely during training to mitigate distribution shift—has been explored in several prior works. For instance, Jones et al. [2024] disentangled style and content learning by first training a content LoRA and subsequently a style LoRA, using only the latter at inference to achieve clean style transfer. Similarly, Guo et al. [2023] fine-tuned a LoRA on the final video dataset to better absorb the target distribution shift. Ostris [2024] proposed a LoRA variant capable of "un-distilling" Flux, allowing users to fine-tune the already step-distilled Flux[schnell] model. Building on these insights, our work generalizes this concept beyond LoRA-based adapters: we demonstrate that auxiliary modules such as ControlNet can likewise be trained to absorb spurious correlations—e.g., those related to pose or expression—thereby isolating shortcut factors from the main adapter's representation.

3 Shortcut-Rerouted Adapter Training

3.1 Mathematical Formalism

We formalize adapter training within a probabilistic framework to clarify the core challenge of disentanglement. Let X denote the observed image, which depends on two underlying factors: the *target factor* T (e.g., identity, style) and the *confounding factors* C (e.g., pose, distribution shift, expression). Formally, we assume:

$$X \sim p(X \mid T, C), \tag{1}$$

where T is the factor we wish to faithfully capture via adapter training, and C represents incidental attributes that are not of primary interest.

The adapter $\mathcal{A}$ is a function that takes X as input and produces a representation $\mathcal{A}(X)$, which is used to steer the generative model G. Ideally, we seek:

$$G(\mathcal{A}(X)) \approx p(\cdot \mid T), \tag{2}$$

meaning that the adapter should extract and inject only the information relevant to T, regardless of the confounds C.

However, in typical single-image reconstruction training (Fig. 2), the objective is to minimize:

$$\mathbb{E}_{(X)}\big[\mathcal{L}\big(G(\mathcal{A}(X)), X\big)\big], \tag{3}$$

which implicitly encourages $\mathcal{A}(X)$ to encode *both* T and C, since X embodies all these factors. As a result, the adapter becomes entangled: instead of isolating the target factor T, it captures spurious correlations mediated by C. These *shortcuts*, by minimizing the objective through confounding factors, inadvertently become reinforced and impair generalization to test prompts.

We propose **Shortcut-Rerouted (SR) Adapter Training**. As demonstrated in Fig. 3, our key insight is to reroute the shortcuts, i.e. the influence of C, through an auxiliary module $\mathcal{S}_C$, which explicitly utilizes the confounding factors. The generative process is modified to:

$$\hat{X} = G(\mathcal{A}(X), \mathcal{S}_C(C)), \tag{4}$$

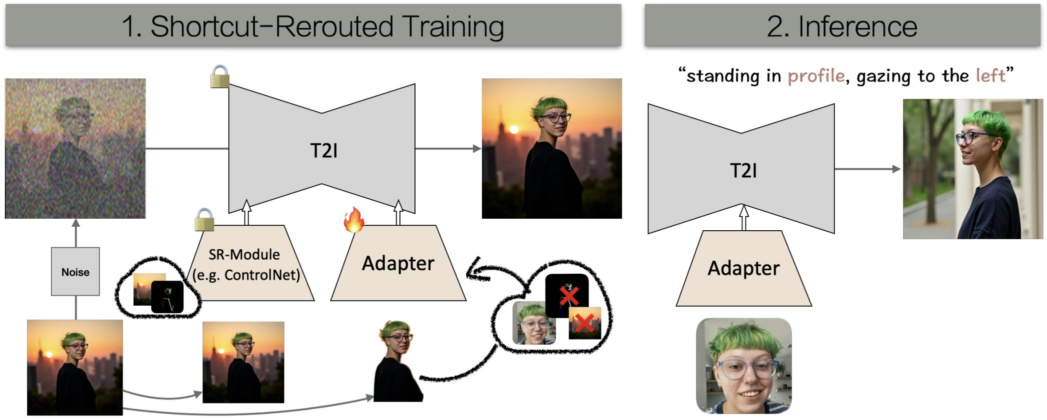

Figure 3: **Method.** The Shortcut-Rerouting (SR)-Module serves as a generic shortcut adapter that can take various forms—such as a ControlNet, LoRA, or IP-Adapter—depending on the confounding factor being addressed (e.g., pose, distribution, or style). Illustrated here is the case of SR with ControlNet, where pose and expression cues are explicitly rerouted via the ControlNet during adapter training. At inference time, the ControlNet is removed, restoring independent pose and expression control from the text prompt alone.

where $\mathcal{S}_C$ is a pre-trained and frozen module that directly provides C to the generator, i.e. establishes the *shortcuts*. The revised training objective becomes:

$$\mathbb{E}_X \left[\mathcal{L}\big(G(\mathcal{A}(X), \mathcal{S}_C(C)), X\big) \right], \tag{5}$$

which ensures that the confounding factors are explained away by $\mathcal{S}_C$, leaving $\mathcal{A}(X)$ with no incentive to encode them. In effect, we turn the entanglement problem into a modular decomposition: $\mathcal{A}$ is pressured to specialize in T, while $\mathcal{S}_C$ accounts for C during training.

Finally, **the shortcut module $\mathcal{S}_C$ is removed during inference**, recovering the original model, but equipped with a disentangled adapter. Now, the generative process in inference, $\hat{X} = G(\mathcal{A}(X))$, is less likely be impacted by the confounding factors from X.

This formulation reflects a general principle: by *explicitly modeling* nuisance factors during training, we prevent the adapter from internalizing them, yielding cleaner and more robust representations.

3.2 Personalized Generation via Shortcut-Rerouted(SR) T2I Adapter Training

Adapter training aims to steer a frozen text-to-image model by injecting additional signals—typically derived from a reference image—into its generation process. In the setting of personalized generation, adapters are lightweight modules that encode the input subject's identity and modulate the diffusion model to personalize its output accordingly.

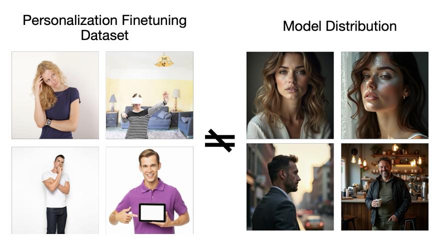

Figure 4: **Distribution shift between model and finetuning dataset.** Due to the distribution shift, directly training a personalization adapter on the finetuning dataset leads to degraded quality.

3.2.1 Instantiations of SR Module

A central challenge in conventional adapter training is that the adapter often encodes confounding factors—such as distribution biases in the fine-tuning dataset or pose and lighting leakage from the input images—thereby entangling the target identity with spurious features. The SR module $\mathcal{S}_C$ in Eq. (4) is versatile, and can be realized by many different modules capable of absorbing specific confounders. While in principle multiple such modules can be composed to form a single unified SR module, in this work we focus on two primary instantiations: SR-LoRA, which addresses dataset-level distribution shifts, and SR-CN, which handles pose and expression leakage.

3.2.2 SR-LoRA: Addressing distribution shift

The first application of Shortcut-Rerouted Adapter Training addresses the issue of *distribution shift* between the model distribution and the data distribution used during adapter finetuning. In many real-world scenarios—particularly with proprietary models such as Flux—the training distribution of the backbone model is unknown or opaque. Meanwhile, personalization pipelines often finetune on curated datasets with specific styles, subjects, or domains. This mismatch introduces a latent confounding factor: *the domain gap between the foundation model and the finetuning data*.

To absorb this domain-induced shortcut, we instantiate $\mathcal{S}_C$ as a light-weight LoRA module, trained specifically to capture this distributional gap. Concretely, we pretrain this LoRA on the finetuning dataset (e.g., studio-lit identity images), allowing it to absorb the dataset-specific style, lighting, and low-level features that differ from the base model's prior. During adapter training, we then freeze the LoRA and train the identity encoder $\mathcal{A}$ as the only active module, allowing it to focus solely on identity, independent of the dataset domain:

$$\hat{X} = G(\mathcal{A}(X), \mathcal{S}_C(C)),$$

where $\mathcal{S}_C$ provides the latent adjustment required to bridge the domain gap, rerouting the shortcut through a controlled path. As in our general formulation, this ensures that $\mathcal{A}(X)$ is no longer incentivized to account for the domain discrepancy, and instead focuses on learning a representation faithful to T.

At inference time, we remove $\mathcal{S}_C$, resulting in a generation pipeline governed solely by $\mathcal{A}(X)$. This yields identity adapters that generalize beyond the specific visual domain of the training data and respond more reliably to test-time prompts across domains. In effect, this use case demonstrates that even abstract or latent confounders—such as dataset shift—can be systematically absorbed via shortcut modules, extending our methodology beyond structured confounds like pose or expression.

3.2.3 SR-CN: Addressing Pose and Expression Leakage

Going beyond absorbing distribution shift, we aim to address the challenge of absorbing the shortcuts of expression and pose from the input image during inference. The target factor T is facial identity, while the target confounding factors C_{CN} for this SR module include head pose and facial expression. To absorb C_{CN}, we employ a pre-trained ControlNet [Zhang et al., 2023] module $\mathcal{S}_{CN}$ that conditions generation on pose and expression maps derived from the training images.

During adapter training, we augment the generative pipeline as follows: given a training image X and its corresponding identity T and pose/expression C_{CN}, we generate pose/expression maps (e.g., via pose estimation and landmark detection) and feed these to $\mathcal{S}_{CN}$. The adapter $\mathcal{A}$ is then trained to inject identity alone, while $\mathcal{S}_{CN}$ accounts for the pose and expression. The overall objective in adapter training stage becomes:

$$\mathcal{L}\big(G(\mathcal{A}(T), \mathcal{S}_{CN}(C_{CN})), X\big) \tag{6}$$

which, as shown in Fig. 1, leads to adapters that are robust across a wide range of poses and expressions. See Fig. 3 for an illustration.

This approach highlights the generality and modularity of our method: by providing a controlled pathway for spurious factors, we relieve the adapter from modeling them, resulting in cleaner and more generalizable representations.

4 Experiments

In this section, we show a number of experiments for different instance of Shortcut-Rerouted Adapter Training. First, we show results for Shortcut Rerouting in the setting of 'face' adapters using both LoRA and ControlNet as Shortcut Rerouting mechanisms (§4.2). The resulting adapters demonstrate improved prior preservation, head pose control and expression control. Then, we show results for Shortcut Rerouting in the setting of 'body' adapters (§4.3).

4.1 Experimental Setup

Datasets. We curate an internal large-scale dataset of a few million high-quality human images, filtered to retain only single-subject photos and remove low-quality, NSFW, or watermarked content. To

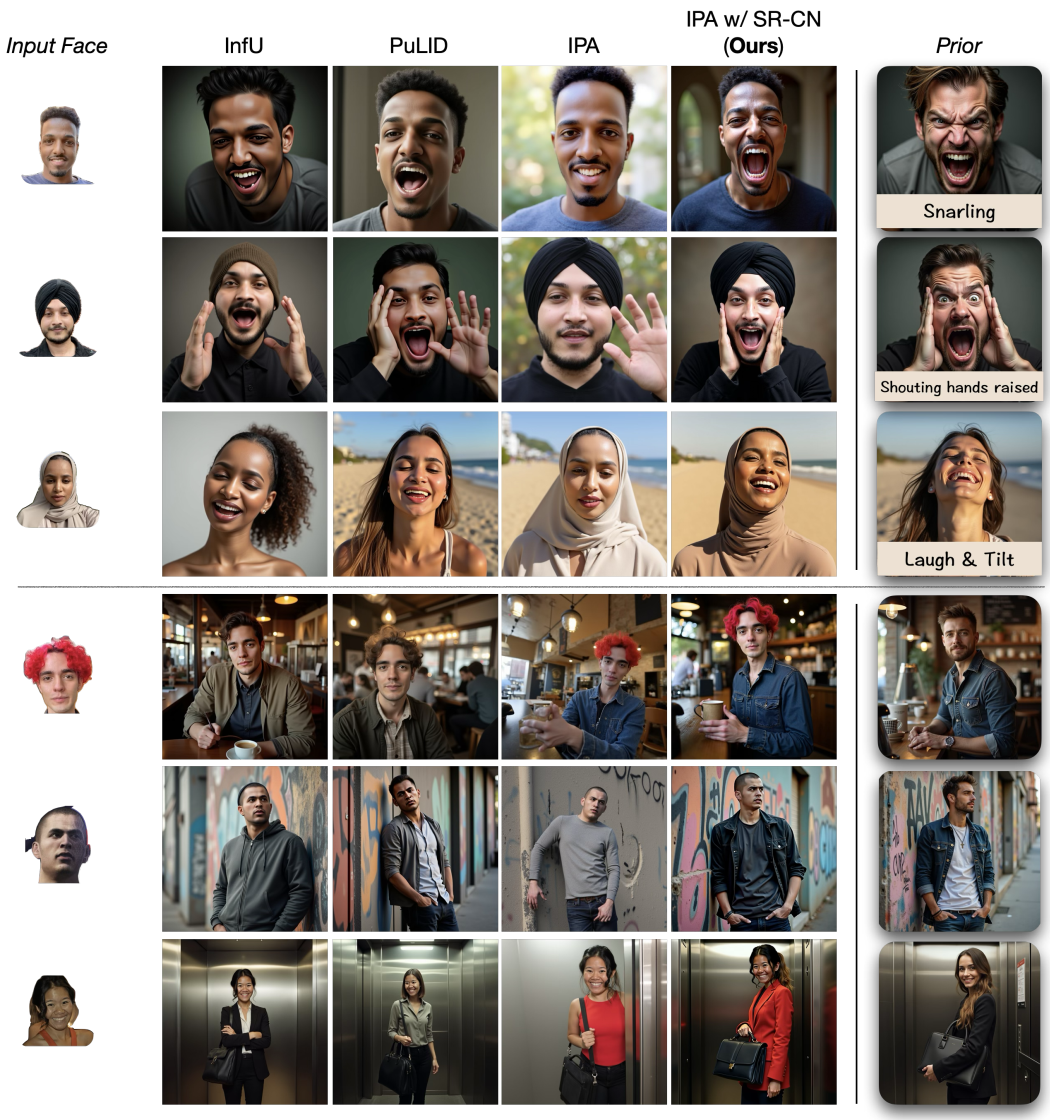

Figure 5: **Qualitative comparison of different "face" adapters.** *Top*: close-up portraits with varied expressions. *Bottom*: full-body generations. Our approach preserves the visual prior more faithfully, enabling expressive and identity-consistent personalized image generation.

accelerate training, we bucket images by aspect ratio and cache auxiliary modalities such as landmarks, segmentation masks, and text embeddings. For adapter inputs, we extract and align face crops using facial landmarks; for full-body, we extract body crops via segmentation and apply background removal. Captions are generated using Qwen2.5-14B (text) and InternViT-300M-V2.5 (vision), both state-of-the-art large-scale captioning models. We also provide the details and visualization for test input images and prompts in the Appendix.

Training and Implementation details. All methods are implemented in `PyTorch` [Paszke et al., 2019] using the `HuggingFace Diffusers` [von Platen et al., 2022] framework, based on the `FLUX.1 [Dev]` [Labs, 2024] model with a DiT [Peebles and Xie, 2023] backbone and Conditional Flow Matching objective [Esser et al., 2024]. Training is performed on $8\times$A100 GPUs (80GB each) using AdamW [Loshchilov and Hutter, 2019] with a learning rate of $5\mathrm{e}{-5}$ and a global batch size of 32 for 250K iterations. Inference is standardized across all methods with IP scale 1.0, CFG 3.5, 28 steps, and 1024×1024 resolution. For identity encoding, we use `openai/clip-vit-large-patch14` [Radford et al., 2021].

Table 1: **Quantitative comparison for "face" adapters.** Our Shortcut-Rerouted methods (SR-LoRA and SR-ControlNet) outperform prior baselines in head pose control and prior preservation, while maintaining competitive identity fidelity. All models are based on Flux Dev.

Method	LLM Id. ↑	FaceNet Id. ↑	LLM Expr. ↑	EMOCA Expr. ↑	Head Pose ↓	Prior (LPIPS) ↓
InfU [Jiang et al., 2025]	3.3824	0.7402	**3.7664**	0.5420	17.7139	0.4490
PuLID [Guo et al., 2024]	4.2826	**0.7742**	3.5899	0.4890	17.5345	0.4584
IPA [Ye et al., 2023]	4.7929	0.7150	3.0714	0.3470	16.1199	0.4800
SR-LoRA IPA (**Ours**)	4.7194	0.6708	3.4286	0.4580	13.2701	0.4330
SR-CN IPA (**Ours**)	**4.7941**	0.7118	3.6934	**0.5800**	**12.6755**	**0.3937**

Table 2: **Quantitative evaluation of "body" personalization methods.** Our SR-ControlNet outperforms both InstantX and baseline IPA across all metrics, showing improved disentanglement and better adherence to pose and expression prompts without sacrificing identity.

Method	LLM Id. ↑	FaceNet Id. ↑	LLM Expr. ↑	EMOCA Expr. ↑	Head Pose ↓	Body Pose ↓	Prior (LPIPS) ↓
InstantX [InstantX, 2024]	2.9930	0.3533	3.4736	0.4687	25.97	186.7454	0.5075
IPA [Ye et al., 2023]	4.5986	0.5733	3.3000	0.3466	20.70	167.4000	0.4566
SR-CN IPA (**Ours**)	**4.6510**	**0.5857**	**3.5263**	**0.4794**	**18.05**	**137.6888**	**0.4133**

Metrics used to measure id preservation, prompt following, and prior preservation are follows:

1. **FaceNet Id.** (↑) – cosine similarity of the FaceNet [Schroff et al., 2015] embeddings of the generated image and the input subject.
2. **LLM Id.** (↑) – a LLM-as-a-judge score for holistic identity. A good metric for personalization should capture the resemblance of the face, head, and hair. While existing Face metrics based on recognition models can capture the cropped face, it does not measure the head/hair. To have a more holistic measure, we use LLM-as-a-judge similar to recent studies [Luo et al., 2024].
3. **LLM Expr.** (↑) – a LLM-as-a-judge score for alignment between the expression specification in the prompt, and the expression in the generated images.
4. **EMOCA Sim.** (↑) – cosine similarity between facial expression embeddings of the generated image and the prior image. The embeddings are extracted using EMOCA model [Daněček et al., 2022].
5. **Head Pose.** (↓) – the mean absolute difference in head orientation (i.e. yaw, pitch, and roll) between the generated and prior images, measured in degrees. We use HopeNet [Ruiz et al., 2018] to estimate these angles.
6. **Body Pose.** (↓) – the mean L2 distance between estimated 2D body keypoints of the generated and prior images in pixel space. HRNet [Sun et al., 2019] is used for keypoint estimation.
7. **Prior (LPIPS)** (↓) – the LPIPS [Zhang et al., 2018] between the generated image and the prior image. Used to measure how much the generated image deviated from the prior.

More details such as the instructions for the LLM-as-a-judge metrics can be found in the Appendix.

Baselines. Our experiments are conducted under two different settings: one where only the *face* is used as input, and another where the fully *body* is provided. Baselines for face-input setting include InfU [Jiang et al., 2025], PuLID [Guo et al., 2024], and an IP Adapter [Ye et al., 2023] trained by us without shortcut rerouting. For the body-input setting, we include an open source model from InstantX, namely `InstantX/FLUX.1-dev-IP-Adapter` [Team, 2024]. This is a general-purpose IPA and not specifically trained for human faces, but serves as a representative baseline for comprehensive evaluation.

4.2 "Face" Adapters

Absorbing distribution shift. Our first key result, shown in Table 1, is the substantial improvement in prior preservation scores for both SR-LoRA and SR-CN. These gains are also clearly reflected in the qualitative examples in Fig. 5, where the image layout, texture, and overall visual quality remain closely aligned with the reference prior. In contrast, all baseline methods—including the IPA variant without shortcut rerouting—exhibit noticeable deviations in texture and scene fidelity. The

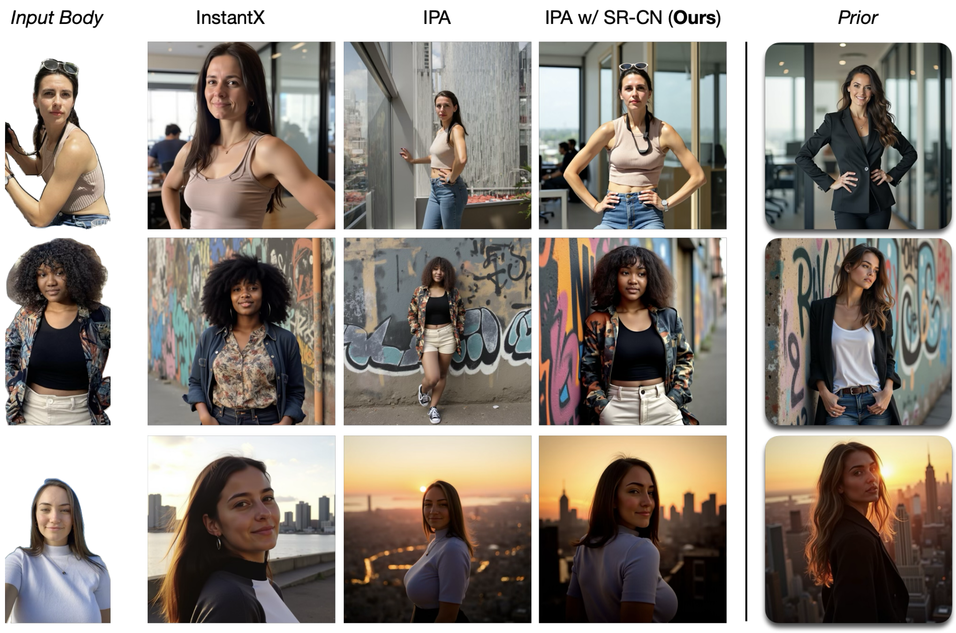

Figure 6: **Qualitative comparison of different "body" adapters.** Our approach shows much stronger identity preservation than InstantX [InstantX, 2024], and much better adherence to the prior and enhanced image quality when compared to vanilla IPA [Ye et al., 2023].

improvement in visual consistency highlights the effectiveness of shortcut rerouting in absorbing distributional differences during training.

Enabling text-guided control of pose and expression. Our second key result is that SR-CN better preserves and generalizes over mutable aspects of a person's identity—such as pose and expression—compared to standard IPA. *Vanilla IPA suffers from pose and expression shortcuts, often copying these attributes directly from the input image.* For example, in the first row of Fig. 5, the subject is smiling in the reference image, causing the output to ignore the prompt-specified expression of "snarling." Other baselines, such as InfU and PuLID, also struggle with identity fidelity, showing noticeable inaccuracies in head shape and hairstyle. It is worth noting that unfortunately face identity distance cannot measure this identity shift caused by head shape and hairstyle. In contrast, the more disentangled adapter trained with shortcut rerouting not only respects prompt-driven expression but also supports more natural and coherent full-body generations (see bottom half of Fig. 5).

4.3 "Body" Adapters

Beyond just face: Capturing identity aspects like body type, clothing, and limb proportions. We explore training adapters using full-body crops as input, which provide a richer signal for capturing holistic identity traits such as body type, clothing, and limb proportions—factors critical for realism and character consistency in downstream generations. *However, this richer input also increases the risk of shortcut learning, particularly the tendency to copy the body pose from the reference image.* As a result, desired applications like reposing a subject through text prompts become more difficult.

As shown in Table 2, IPA trained with shortcut rerouting achieves the highest performance across all metrics: identity fidelity, pose controllability, and prior adherence. These improvements are clearly visible in Fig. 6, where SR-CN IPA not only preserves subject identity more faithfully, but also *produces outputs that are more consistent with the prior image layout and appearance*.

4.4 "Background" Adapter

Additional variants that include further modules (e.g., background adapters) and extended combinations such as SR-LoRA+CN, and SR-LoRA+CN+BG are discussed in the Appendix, along with ablation results illustrating their complementary effects. As an example depicted in Fig. 7, the LoRA

Figure 7: **SR-Training is a versatile framework supporting different combinations of shortcut modules**. The LoRA shortcut mitigates quality degradation, producing generations more consistent with the prior compared to the baseline. The ControlNet (CN) shortcut preserves pose priors, while the background (BG) shortcut prevents lighting leakage from the input. Notably, SR-LoRA-CN follows the pose of the prior but deviates in the background, whereas SR-LoRA-BG preserves the background but deviates slightly in pose. Finally, SR-LoRA-CN-BG aligns closely with both pose and background, thereby isolating and injecting only the target identity.

shortcut substantially alleviates quality degradation, the ControlNet (CN) shortcut reliably maintains pose priors, and the background (BG) shortcut effectively suppresses illumination leakage.

Limitations. In this work, we introduced Shortcut-Rerouted Adapter Training, a simple yet broadly applicable framework for disentangling spurious correlations in text-to-image personalization. One limitation with the current evaluation is our focus on encoder-based adapter training. In principle, our approach could also be applicable to LoRA training, such as learning a style LoRA free of undesired shifts like layout, or content. Specifically for the task of personalization, one limitation is the fact that we applied Shortcut-Rerouting to IP-Adapter only, which is a fairly simple baseline. Applying our approach to stronger baselines might lead to overall better performance.

Ethical Considerations. As our method improves identity preservation and expression control in personalized generation, it naturally raises concerns about misuse, particularly in the creation of hyper-realistic synthetic identities or deepfakes. We acknowledge that enhanced controllability and realism may lower the barrier for malicious use. To mitigate such risks, we advocate for responsible deployment practices, such as model watermarking, usage restrictions, and alignment with ethical frameworks for generative media.

5 Conclusion

We introduced Shortcut-Rerouting, a simple yet general framework for disentangling spurious correlations in text-to-image personalization. By explicitly providing shortcut pathways for confounding factors during training—via modules such as LoRA or ControlNet—we prevent adapters from internalizing undesirable attributes like pose, expression, or domain-specific biases. This leads to cleaner, more controllable representations that preserve identity while restoring the model's ability to respond to prompt-based instructions. Our experiments across both face and full-body personalization demonstrate improved controllability, prior preservation, and generation quality.

More broadly, our results suggest a general principle: when training models to focus on what matters, it is often most effective to explicitly route away what does not. We believe this perspective of Shortcut Rerouting has implications beyond personalization: the idea of absorbing confounding variation through targeted pathways can inform future approaches to modular, interpretable, and more controllable generative systems.

References

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, oct 2022. doi: 10.1109/tpami.2021.3087709. URL https://doi.org/10.1109%2Ftpami.2021.3087709.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*. OpenReview.net, 2024.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*. OpenReview.net, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning ID customization via contrastive alignment. *CoRR*, abs/2404.16022, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

InstantX. FLUX.1-dev-IP-Adapter. https://huggingface.co/InstantX/FLUX.1-dev-IP-Adapter, 2024. Accessed: 2024-11-01.

Md. Amirul Islam, Sen Jia, and Neil D. B. Bruce. How much position information do convolutional neural networks encode? In *ICLR*. OpenReview.net, 2020.

Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025.

Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–13, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E Gonzalez, Ruslan Salakhutdinov, Ion Stoica, et al. Stylus: Automatic adapter selection for diffusion models. *Advances in Neural Information Processing Systems*, 37:32888–32915, 2024.

Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34:13073–13085, 2021.

OpenAI. Gpt-4o technical report, 2024. URL https://openai.com/index/gpt-4o. Accessed: 2025-05-22.

Ostris. Flux.1-schnell-training-adapter, 2024. URL https://huggingface.co/ostris/FLUX.1-schnell-training-adapter. Accessed: 2025-10-15.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182. IEEE, 2023.

Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Omni-id: Holistic identity representation designed for generative tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8786–8795, 2025a.

Guocheng Gordon Qian, Daniil Ostashev, Egor Nemchinov, Avihay Assouline, Sergey Tulyakov, Kuan-Chieh Jackson Wang, and Kfir Aberman. Composeme: Attribute-specific image prompts for controllable human image generation. *arXiv preprint arXiv:2509.18092*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 815–823, 2015.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

InstantX Team. Instantx flux.1-dev ip-adapter page. https://huggingface.co/InstantX/FLUX.1-dev-IP-Adapter, 2024.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A single-pass lora-free model for stylized image generation. *ArXiv*, abs/2309.01770, 2023. URL https://api.semanticscholar.org/CorpusID:261531689.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

6 Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: At the time of submission, we are unable to release the code or dataset due to legal and compliance constraints within our organization. We recognize the value of open access and reproducibility and are actively exploring the possibility of releasing portions of the code or evaluation tools pending internal review. We provide detailed descriptions of our methodology, datasets, and evaluation setup in the paper and appendix to support reproducibility in principle.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Comparing LLM-as-a-judge with FaceNet as a metric for Identity Preservation

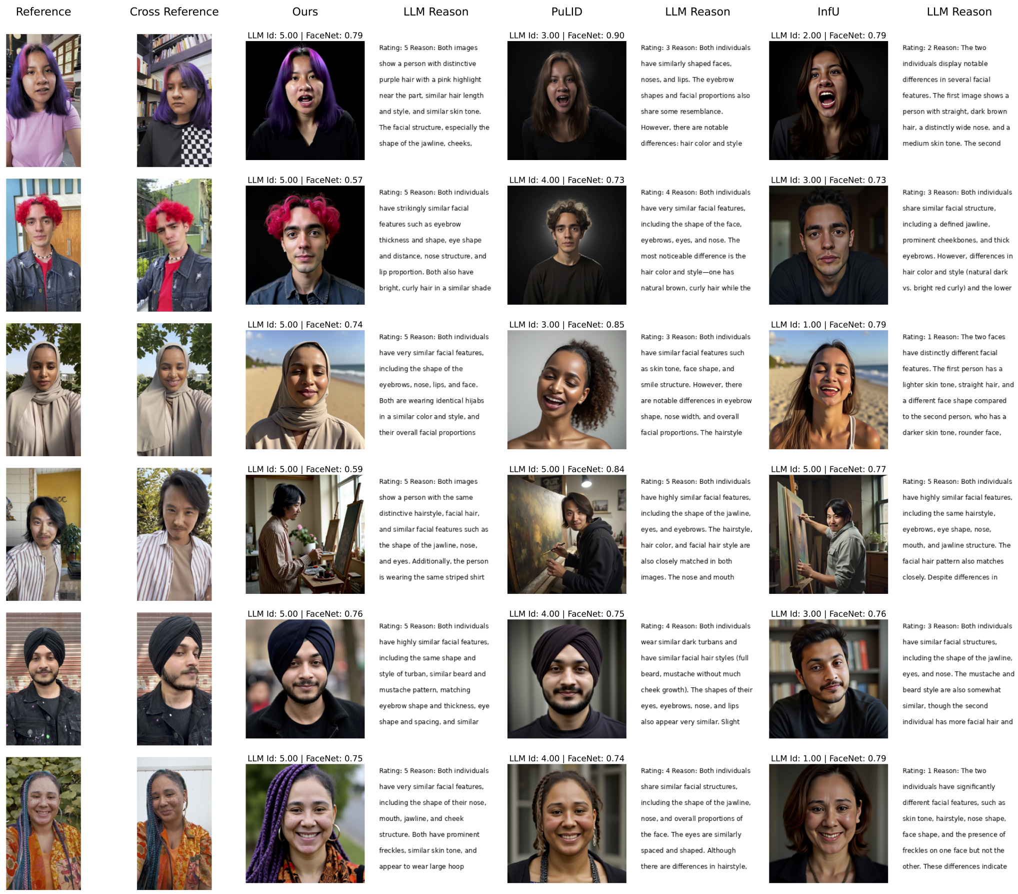

Figure 8: Qualitative comparison between LLM-based identity score and FaceNet Similarity. The face crop from the reference image is used as input to the face adapter. The cross reference image is a shifted view of the same person and is used to compute identity scores.

We observe that face recognition models such as FaceNet [Schroff et al., 2015] and ArcFace [Deng et al., 2022], commonly used for face similarity metrics, fail to account for variations in hairstyle, hair color, headwear, and head shape. We hypothesize that this limitation stems from their training setup, where the input face is tightly cropped to the size of $0.8\times$ of FFHQ [Gal et al., 2021]), resulting in a narrow understanding of facial identity. Consequently, these models provide unreliable assessments of identity similarity in more holistic capture including hairstyle. To address this, we leverage large language models (LLMs) with strong multimodal capabilities—specifically ChatGPT-4o [OpenAI, 2024]—to perform identity assessments. As illustrated in Fig. 8, the LLM accurately captures identity similarity in cases where FaceNet or ArcFace fails.

B LLM-Based Identity Evaluation Prompt

We use an LLM-as-a-judge to evaluate identity preservation in generated images. The LLM is given two face images and asked to judge whether they show the same person. The prompt is shown below:

> **You are an expert in facial recognition. Given two images of faces, your task is to judge whether they show the same person.**
> **Steps:**

1. Carefully compare facial features like shape, eyes, nose, mouth, jawline, eyebrows, and overall proportions.
2. Ignore lighting, angle, pose, worn accessories, or expressions.
3. Rate identity similarity on a scale from 1 to 5:
   - 5 = Definitely the same person
   - 4 = Very likely the same person
   - 3 = Possibly the same person
   - 2 = Unlikely the same person
   - 1 = Definitely not the same person
4. Explain your reasoning by referencing specific facial features.

Respond in the following format:

Rating: <1–5>

Reason: <explanation>

C LLM-Based Expression Evaluation Prompt

We use an LLM-as-a-judge to assess how well generated images reflect the intended facial expressions described in text prompts. The model is instructed to evaluate *only* the facial expression, disregarding pose, clothing, background, or lighting. The exact prompt is provided below:

You are an expert judge tasked with evaluating how accurately a generated image captures the facial expression described in a text prompt. Your evaluation should focus *only on the expression* (such as happiness, sadness, anger, surprise, etc.) and ignore other factors like the person's pose, clothing, background, or lighting.

Steps:

1. Carefully read the text prompt.
2. Examine the facial expression in the generated image.
3. Compare the image's expression to the description in the prompt.
4. Rate the **expression fidelity** on a scale of 1 to 5:
   - 5 = Perfect match.
   - 4 = Mostly accurate with minor discrepancies.
   - 3 = Partially matches but has clear inaccuracies.
   - 2 = Mostly incorrect expression.
   - 1 = Completely wrong expression.
5. Briefly explain **why** you gave this rating, pointing out specific facial features (mouth, eyes, brows, etc.) that contributed to your assessment.

Respond in the following format:

Rating: <1–5>

Reason: <explanation >

D Enabling expression, pose, and lighting control

In this section, we highlight the fine-grained control over pose, expression, and lighting enabled by our shortcut-rerouted training. In 9, pose varies across columns while holding expression constant, and expression/lighting vary across rows with fixed pose. This disentangled control is achieved while faithfully preserving identity from the reference image.

E Additional Qualitative Results for Face Adapters

F Additional Qualitative Results for Fullbody Adapters

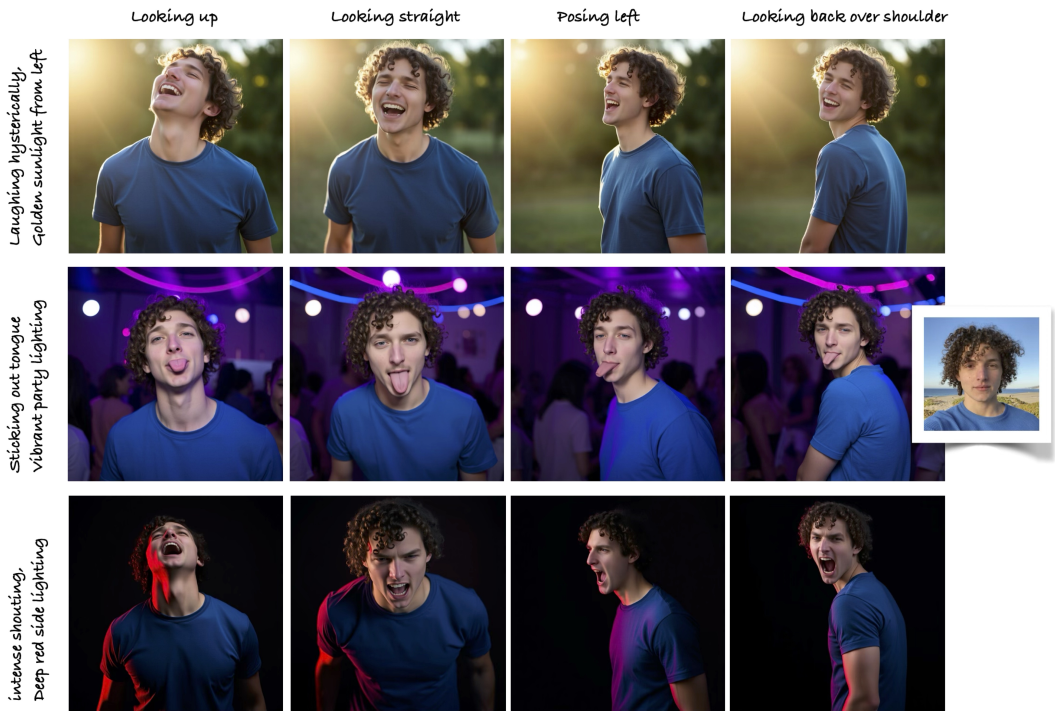

Figure 9: **Shortcut Rerouting enables to precisely control lighting, pose, and expression from text.** Adapter preservers from prior while utilizing only personalization related attributed from the reference image. Zoom in for best view.

Figure 10: **Qualitative comparison of different "face" adapters.** We evaluate adapters on their ability to edit and preserve a wide variety of facial expressions, poses, accessories, lighting, and activities. **SR-LoRA** addresses quality degradation; **SR-CN** achieves better pose control; **SR-LoRA-BG** improves background and lighting consistency; and using both SR-LoRA and SR-CN, namely **SR-LoRA-CN** effectively maintains pose while maintaining the quality; Zoom in for best view.

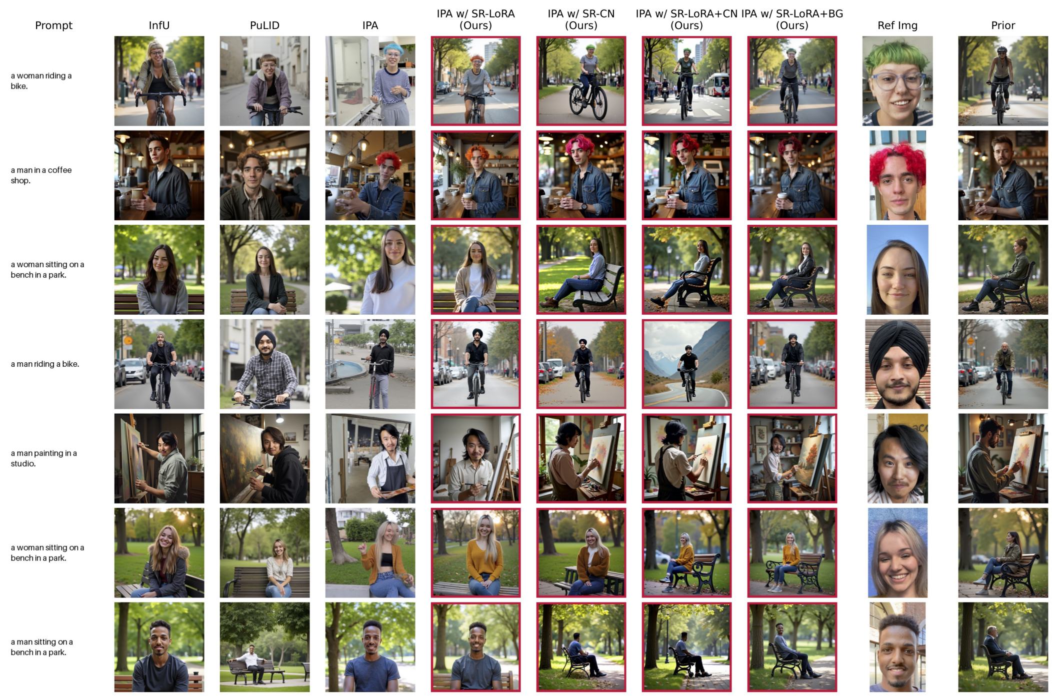

Figure 11: **Qualitative comparison of different "face" adapters.** We evaluate adapters on their ability to edit and preserve a wide variety of facial expressions, poses, accessories, lighting, and activities. **SR-LoRA** addresses quality degradation; **SR-CN** achieves better pose control; **SR-LoRA-BG** improves background and lighting consistency; and using both SR-LoRA and SR-CN, namely **SR-LoRA-CN** effectively maintains pose while maintaining the quality; Zoom in for best view.

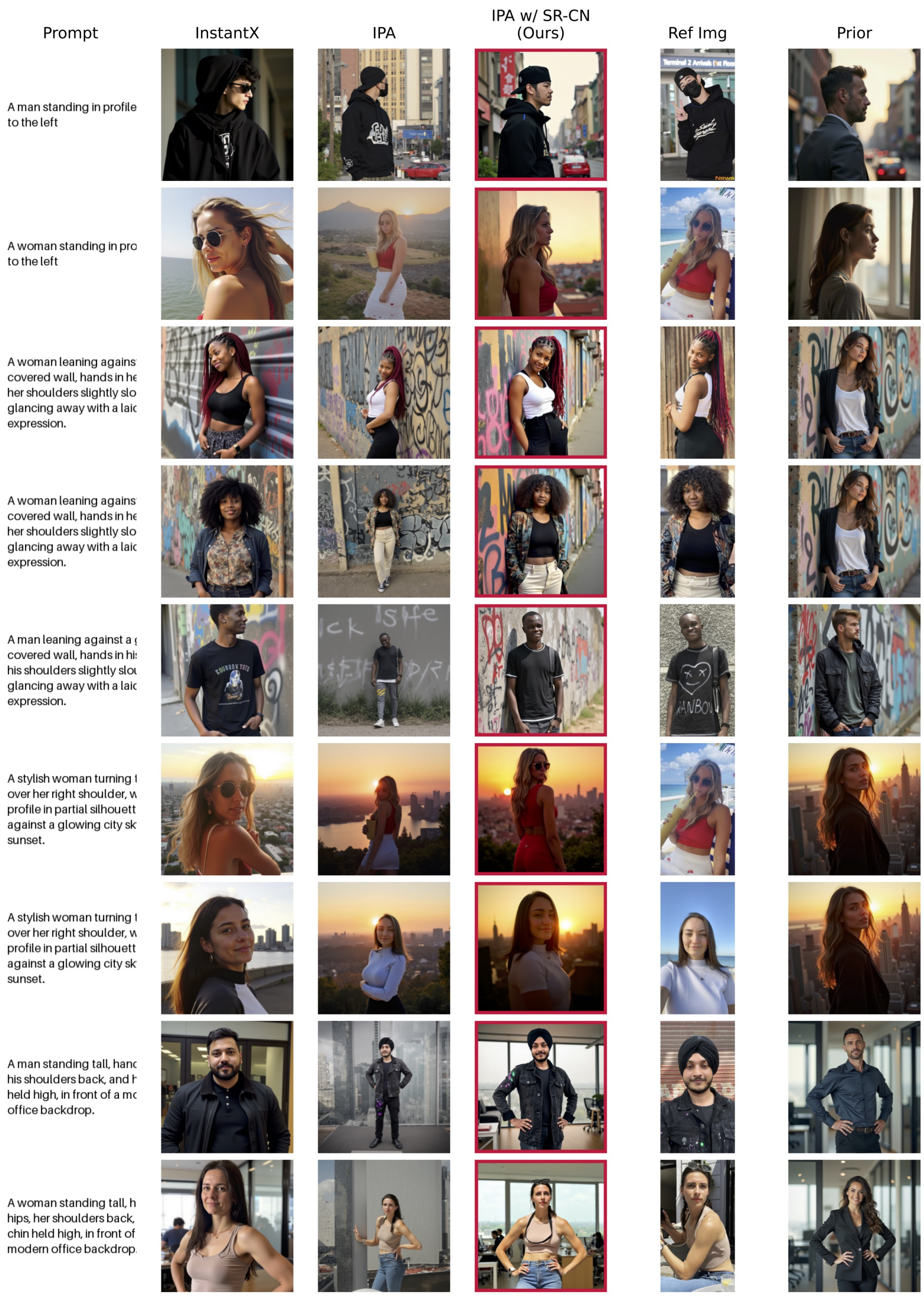

Figure 12: **Qualitative comparison of different "fullbody" adapters.** Our approach consistently outperforms others by more accurately preserving body shape and enabling superior pose edit-ability.

G Evaluation Prompts

We randomly sample four identity images for each of the text prompts below, where 0 serves as a placeholder for the gender noun (e.g., man, woman).

Accessory
A distinguished {0} wearing an ornate, gold-trimmed monocle, paired with a classic black suit, where the monocle is illuminated under bright studio lights.
A fashionable {0} posing under a summer sun, wearing a wide-brim straw hat adorned with fresh flowers, and the camera focusing on the hat's elaborate details.
A glamorous {0} entertainer on stage, wearing extra-long, feathered earrings, with dramatic lighting accentuating their vivid colors and texture.
A glamorous {0} entertainer on stage, wearing extra-long, feathered earrings, with dramatic lighting accentuating their vivid colors and texture.
A modern {0} bride wearing an unconventional black lace veil, highlighted by soft backlighting that emphasizes the veil's intricate pattern.
A {0} striking a pose in an alley, wearing chunky, studded wristbands as the focal point against graffiti-covered walls.
A {0} walking through a busy city street, wearing bold, oversized sunglasses with mirrored lenses that reflect the urban skyline.
A retro-chic {0} fashion model in a vintage polka-dot scarf and matching gloves, with the patterned accessories taking center stage.
A rugged {0} adventurer standing on a cliff, sporting a high-tech, reflective helmet with built-in goggles, making the helmet the centerpiece of the shot.
A stylish {0} dancing in neon-lit surroundings, wearing knee-high lace-up boots sparkling with sequins, with the boots dominating the frame.
An edgy {0} street performer showcasing multiple silver piercings and a spiked choker, with close-up angles capturing every metallic detail.

Activity
A {0} in a coffee shop.
A {0} in the office.
A {0} painting in a studio.
A {0} playing basketball in the NBA.
A {0} riding a bike.
A {0} sitting on a bench in a park.

Expression
{0} biting lower lip nervously, slight frown, brows knit, awkward tension in the face
{0} blowing a kiss, lips puckered, eyebrows raised gently, soft lighting
{0} caught mid-expression between a laugh and a cry, watery eyes, twisted smile, bittersweet emotion
{0} crying intensely, eyebrows arched upward, mouth twisted in pain, eyes squeezed shut, raw emotion
{0} crying with one eye shut tighter than the other, mouth open mid-sob, flushed cheeks
{0} laughing hysterically with eyes shut tight and mouth wide open, cheeks raised, expressive lighting
{0} laughing with head tilted back, eyes closed, mouth wide open, pure joy on the face
{0} roaring in anger, brows furrowed, nostrils flared, teeth bared, dramatic shadowing
{0} screaming with eyes wide and jaw fully dropped, intense emotion on the face, dramatic lighting
{0} shocked with uneven brows, wide eyes, jaw slack, vivid emotion in facial pose
{0} shouting loudly, mouth wide, eyes intense, hands near face, motion blur
{0} smiling with eyes squeezed shut, mouth open in pure elation, cheeks lifted high
{0} smirking with head tilted slightly, one eyebrow raised, subtle attitude in the eyes
{0} snarling with clenched teeth, nose scrunched, eyes narrowed in aggression
{0} sneering in disgust, nose wrinkled, upper lip curled, one eye slightly squinted, gritty atmosphere
{0} sticking out tongue in a mocking expression, playful eyes, raised eyebrow, casual setting
{0} stunned with jaw dropped, eyebrows raised high, eyes wide open, sharp lighting
{0} wincing in pain with eyes tightly shut, lips twisted, brow tense, expressive emotion
{0} winking with a mischievous grin, one eye squinting and shut, playful mood
{0} winking with a mischievous grin, one eye squinting, playful mood
{0} yawning deeply, mouth stretched wide, eyes half-closed, mid-motion, natural lighting

Lighting
A {0} dancing under shifting multicolor disco lights
A {0} gazing sideways, face partially lit by peach morning light
A {0} in a dynamic pose under cold cyan lighting
A {0} in crouched pose under deep indigo overhead spotlight
A {0} leaning against a wall, lit from below with greenish hue
A {0} leaning forward into cool violet side light
A {0} mid-motion, under red and purple colored strobe lights
A {0} reaching upward under golden sunrise rays
A {0} reclining on couch under dusty rose lighting
A {0} sitting cross-legged, lit with emerald green hue
A {0} sitting sideways under diffused teal lighting
A {0} standing tall, shadow cast long under low amber light
A {0} standing under soft orange sunset light, profile facing right
A {0} turning back toward camera, lit from behind with white light
A {0} under warm candlelight, hands resting on lap
A {0} walking into a warm golden spotlight from the side
A {0} with arms crossed, under vibrant magenta rim lighting
A {0} with hands on hips, under bright orange studio lights
A {0} with head tilted back, lit by soft lavender haze
A {0} with one hand in pocket, standing in blue-tinted window light

Pose
A {0} gazing up at a towering waterfall, their head tilted back and arms outstretched
A {0} kneeling on one knee in a dramatic, torchlit cave, leaning forward with eyes fixed on the horizon, exuding focus and intensity.
A {0} leaning against a graffiti-covered wall, hands in pockets, shoulders slightly slouched, glancing away with a laid-back expression.
A {0} seated cross-legged on a wooden floor, hands resting on knees, gazing slightly downward in meditative calm.
A {0} standing in profile, they are gazing to the left
A {0} standing in profile, they are gazing to the left
A {0} standing tall, their hands on their hips, and their chin held high, in front of a modern office backdrop.
A {0} standing tall, with their hands on their hips, and their chin held high, in front of a modern office backdrop.
A stylish {0} turning to glance over their right shoulder, with their profile in partial silhouette against a glowing city skyline at sunset.
A confident {0} facing forward, looking directly into the camera lens with a poised, straight-backed posture, set against a clean studio background.