Preserving Image Properties Through Initializations in Diffusion Models - Supplementary Materials

Jeffrey Zhang jeff@revery.ai
Shao-Yu Chang shaoyuc3@illinois.edu
Kedan Li kedan@revery.ai
David Forsyth daf@illinois.edu

1. Freeform Text Test Set

For our test dataset, we asked three fashion designers to provide eight freeform descriptions of garments to test the generalizability of our text-to-image model. These are completely fictional garment descriptions. The 24 freeform texts we used for testing are:

- (F* or F†) sheer pale pink organza bow collar bomber-style jacket with lace detailed edges, oversized puff sleeves, bow accent pockets, and a lace sailor collar
- (F* or F†) short sheer white dress with white shirt collar
- (F* or F†) baggy blue jeans with red stitching and silver buttons
- (M* or M†) blue shirt with white stripes gold buttons and mao collar
- (F* or F†) asymmetrical blue velvet dress with crochet white pocket
- (F* or F†) white sleeveless dress with ocean blue lining at the bottom
- (F* or F†) black sleeveless dress with ocean blue lining at the bottom
- (F* or F†) black hoodie with white embroidered flowers sleeves
- (F* or F†) black cotton flared pants with sheer dark grey layer
- (F* or F†) low waisted dark wash jeans with distressed knees and flared bottoms
- (M* or M†) brown leather moto jacket with silver buttons and zipper and lace trim from the bottom of the jacket
- (F* or F†) sage green corset top with silk ribbon and no sleeves with a cropped length
- (F* or F†) black ruffled midi skirt with black lace trim along the seam
- (F* or F†) tailored dark wash denim zip corset with pronounced seams and obi inspired bows at the hip
- (F* or F†) silk long tailored straight leg trousers in dark pink with matching ostrich feathers at the hem
- (F* or F†) lilac ostrich leather pleated micro mini skirt with a matte finish and beaded silver details
- (F* or F†) mock black neck cashmere sleeveless top with a heart cutout at the chest and silver threads throughout
- (F* or F†) sheer turquoise silk gown with deep V neckline and train hem, with ruffles on the sleeves and neckline and a cinched tie waist belt
- (F* or F†) leopard print chiffon column gown with a mock neck halter closure, no sleeves, and symmetrical hip cut outs
- (F* or F†) royal blue velvet puffer jacket with an oversized silhouette, stand up scarf collar, silver hardware, and extra long extended sleeves
- (F* or F†) military jacket with clean cut tailoring in dark emerald brocade fabric with crimson red piping and gold accents
- (F* or F†) silver cyber edgy style metal-look top with sleek tribal design that wraps around the body and tie in back with a silver strap to keep it on
- (F* or F†) patchwork pieced together handkerchief skirt with a bohemian style that is maxi length and looks hand-sewn, in many different prints and shades of green fabrics
- (F* or F†) asymmetrical cut denim wrap skirt in a black faded wash. Silver hardware inspired by Chrome Hearts. Rough unfinished hem and one side is longer on the wrap

We use M*/F* to shorthand the description "male/female garment, no person, white background" and we use M†/F† to shorthand the description "male/female person wearing garment." Whenever we aim to generate a garment, we use either M* or F*; whenever we aim to generate a model, we use M† and F†. **All images for these text prompts are displayed in order from left to right and top to bottom, unless mentioned otherwise in the figure.**

Linear End Parameter	α_T
0.012	0.0046600
0.02	0.0002300
0.025	0.00003594
0.03	0.00000567

Table 1. We show linear end values and each corresponding α_T and $\sqrt{\alpha_T}$. Standard training procedures use linear end 0.012.

2. Lowering α Values

Compared to standard training parameters, lowering α values improves the stability of certain properties, but outputs still exhibit inconsistencies in preserving image properties and are often lower quality and underexposed.

We test different values of α_T by adjusting the linear end parameter. Stable Diffusion [1] uses a linear start of 0.00085 and a linear end of 0.012. We keep the same linear start for all experiments and adjust the linear end to lower the magnitude of α_T. We keep all other parameters identical to our finetuning training procedure in the main text. Table 1 shows the altered linear end values and each corresponding α_T value.

Fig. 1 and Fig. 2 show results for garments and models from fine-tuning sd v1.5 [1] on our dataset with different α_T values. Though decreasing α_T brings us closer to our preferred image properties, these properties aren't consistently produced, and the image quality is significantly compromised.

The best garment outcomes are shown at $\alpha_T = 2.30 \times 10^{-4}$, but it still fails to consistently generate uniform white backgrounds (see top right blue mao collar shirt and bottom right patchwork skirt in Fig. 1) and model outputs are still unpredictable. Lowering α_T to 3.59×10^{-5} and 5.67×10^{-6} generates models that are more consistently centered in the image, but images for both garments and models are underexposed with substantial degradation in quality (compare with our Mean Offset Training 50k iteration results in Fig. 8 and Fig. 9).

Small α values create a much harder denoising problem because the amount of noise added is significantly more. Additionally, we are scaling α_t to lower values for many timesteps t, and pre-trained sd v1.5 [1] is not trained to handle large amounts of noise in these timesteps. This results in the need to re-train a greater number of weights and may require a larger dataset to avoid deterioration in image quality.

3. PCA-K Inference on Stable Diffusion v1.5

Applying PCA-K Inference on pretrained Stable Diffusion v1.5 (sd v1.5) [1] can shift the outputted image distribution closer to our desired properties *without* any fine-tuning. In Fig. 3 and Fig. 4, we show how sd v1.5's results change with PCA-3 Inference. With standard noise initialization, we see that none of the images for garments or models are acceptable retail images. However, using PCA-K Inference, the outputted image distribution is closer to satisfying our desired properties, and some may even be acceptable as retail images.

4. Additional Materials for PCA-K Inference and PCA-K Training

4.1. PCA-K Inference Ablations

In general, PCA is a reasonable and easy-to-sample approximation for $P(images)$ because the denoiser behaves similarly between a real image and a PCA-projected image. We show different garment images used as initialization during inference and how results change when these images are projected to different numbers of principal components. We use the standard fine-tuned model on our dataset from our main text (i.e., DDIM training). We use a pink dress (Fig. 5), gray sweater (Fig. 6), and a black long-sleeve (Fig. 7) as reference.

Lighter contrast garments (the pink dress and gray sweater in Fig. 5 and Fig. 6) tend to behave similarly with as little as seven principal components. However, the black sweater (Fig. 7) provides much higher contrast, and the denoiser tends to pick up more information from the initialization. Notice that the sleeves of the projected image are much darker than the center of the garment in PCA-3, PCA-5, PCA-7, and PCA-10. This causes generated results in columns 4, 5, 6, 8, and 9 to be darker in those sleeve regions and lighter in the center regions. Once we use a projection with a darker middle region (e.g., in PCA-20), the generated garments look similar to the original garment initialization result.

4.2. Mean Offset Training

We visualize Mean Offset Training results for all 24 text prompts and show our method can converge to a stable model within 50k iterations. Training parameters are identical to our fine-tuning procedure in the main text. In Fig. 8 and Fig. 9, we compare 10k, 30k, 50k, and 70k training iteration results. Notice that generated garment images satisfy all image properties (1)-(5) within 50k iterations and generated model images satisfy these properties even faster (within 30k iterations).

4.3. PCA-K Training For $K > 0$

PCA-K Training for $K > 0$ exhibits issues during inference due to a disconnect between the initialization and the text. In Fig. 10, we show results from standard fine-tuning and compare results to PCA-1 Training + Inference (Fig. 11), PCA-3 Training + Inference (Fig. 12), and PCA-10 Training + Inference (Fig. 13). For PCA-K Training

+ Inference ($K > 0$), the denoiser still utilizes color and shape information from the initialization and does not fully respect the text (e.g. black sweater initialization causes all the garments to be dark even if the text indicates lighter colors). We leave improving PCA-K ($K > 0$) to future work. One suggestion is to use Canonical Correlation Analysis (CCA) to build relationships between PCA-K initializations and text features. This could help the network sample a smarter PCA-K initialization whose features correlate better with the text.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

Figure 1. Comparison between fine-tuning on different α_T values for garment generation. We use standard noise initialization during inference.

Figure 2. Comparison between fine-tuning on different α_T values for model generation. We use standard noise initialization during inference.

Figure 3. PCA-K Inference (K=3) results for garments on Stable Diffusion v1.5.

Figure 4. PCA-K Inference (K=3) results for models on Stable Diffusion v1.5.

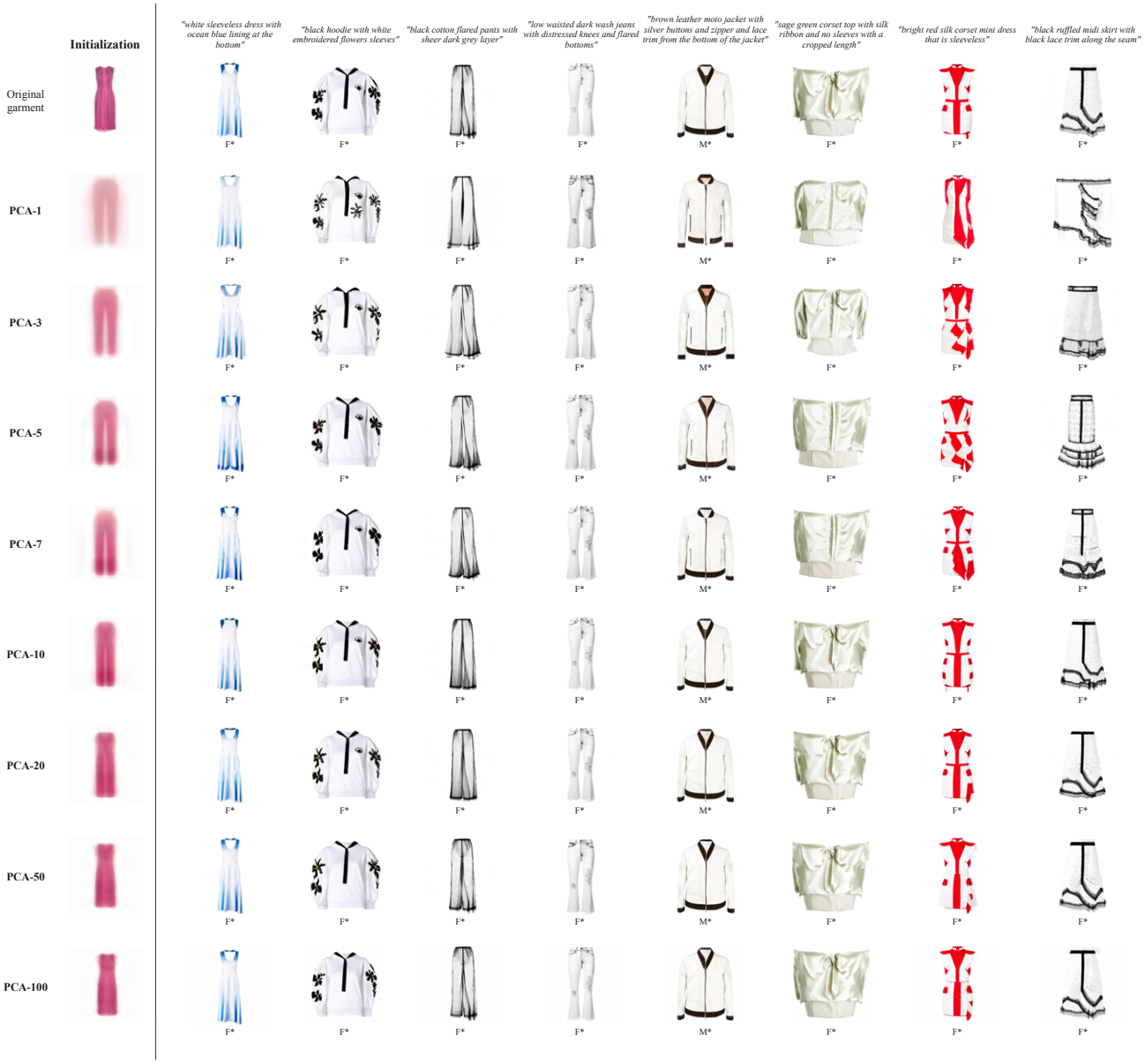

Figure 5. DDIM Training + PCA-K Inference at varying K values for a pink dress.

Figure 6. DDIM Training + PCA-K Inference at varying K values for a gray sweater.

Figure 7. DDIM Training + PCA-K Inference at varying K values for a black long-sleeve.

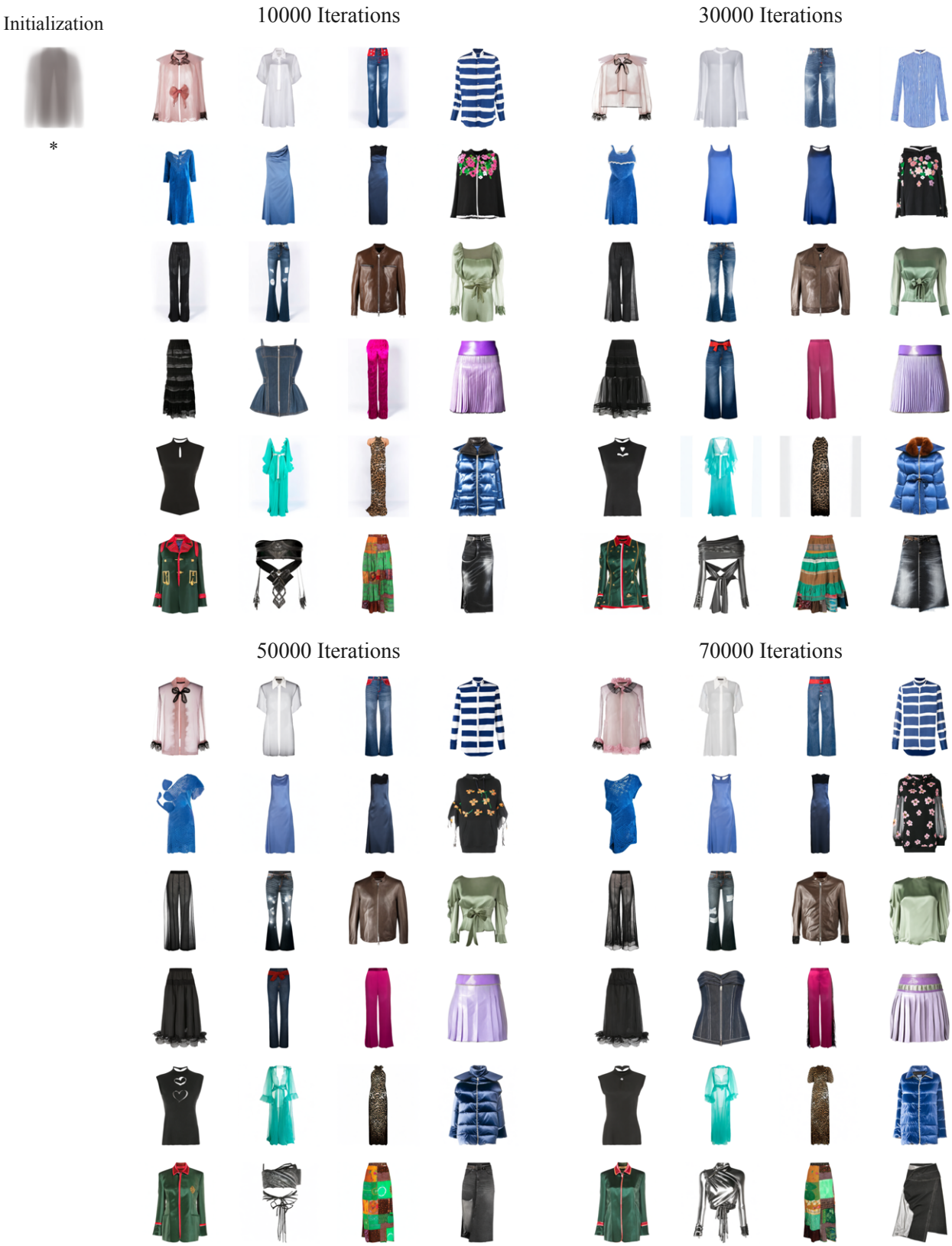

Figure 8. Comparison between training iterations on garment generation for Mean Offset Training + Inference.

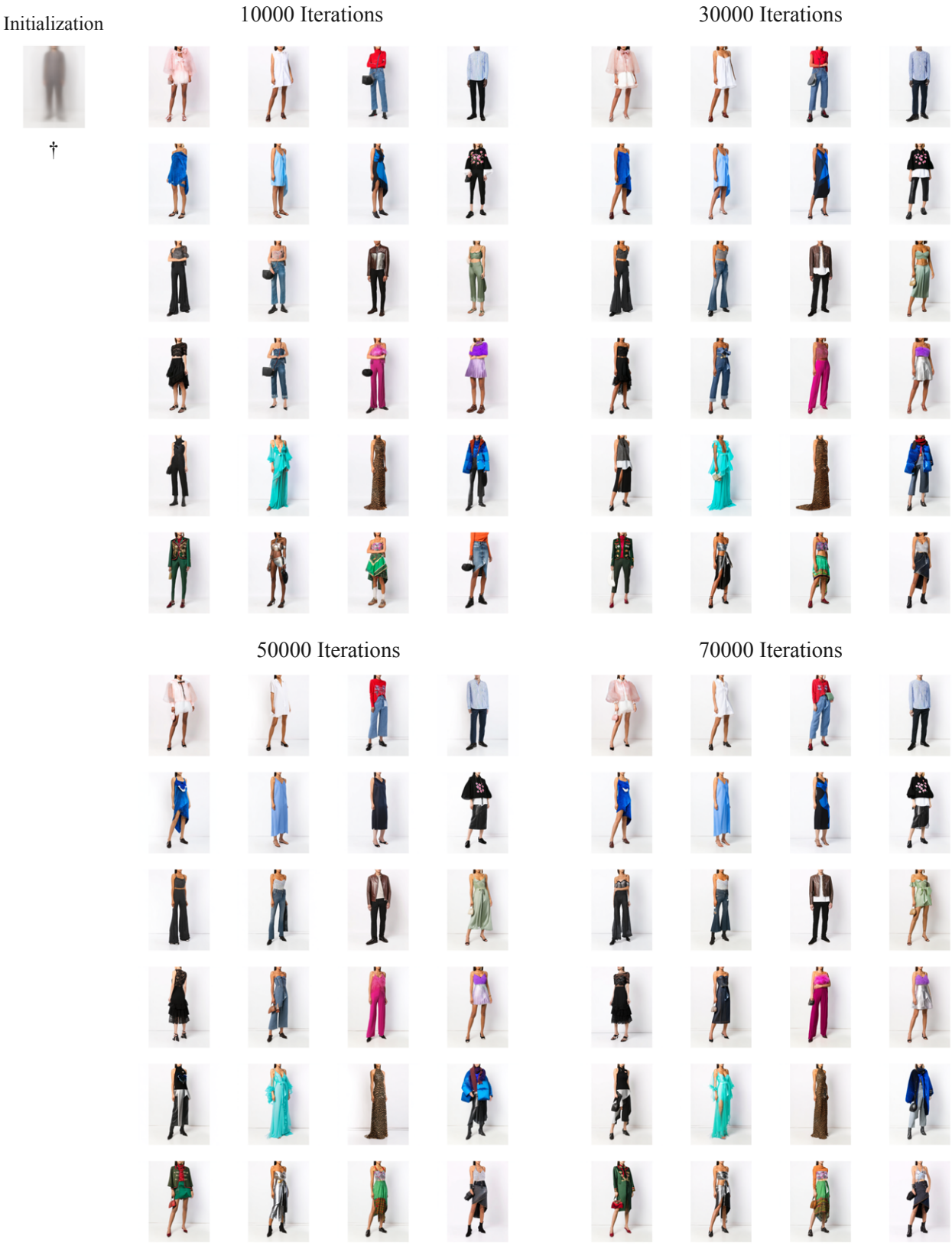

Figure 9. Comparison between training iterations on model generation for Mean Offset Training + Inference.

Figure 10. Comparison results between different initialization with standard training (DDIM Training + Inference).

Figure 11. Comparison results between different initialization with PCA-1 Training + Inference.

Figure 12. Comparison results between different initialization with PCA-3 Training + Inference.

Figure 13. Comparison results between different initialization with PCA-10 Training + Inference.