Appendix: ViscoNet: Bridging and Harmonizing Visual and Textual Conditioning for ControlNet

Anonymous ECCV 2024 Submission

Paper ID #3

This appendix is split into 3 sections:

- **Section A** provides a further quantitative and qualitative comparison with IP-Adatper in the re-identification task (Section 4.2).
- **Section B** showcases more image examples produced by our methods in a variety of tasks, including re-identification, pose re-target, fashion virtual try-on, and stylization.
- **Section C** includes the quantitative results from the main paper with further analysis (Section 4.1, 4.3).

A Re-identification: Comparing IP-Adapter

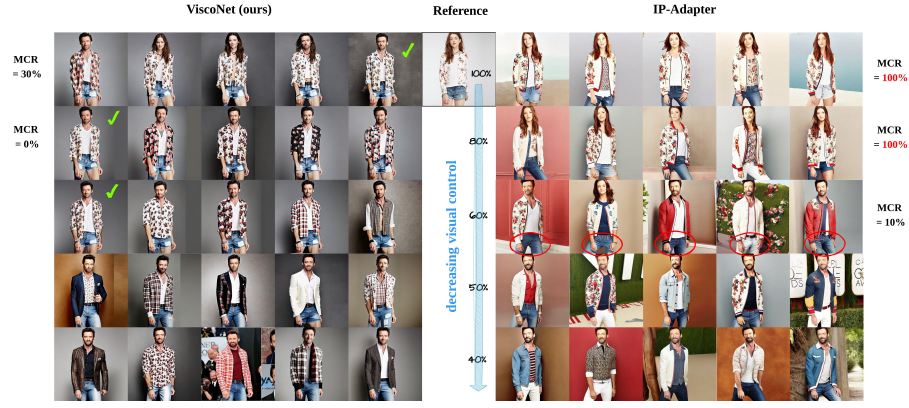
In the Section 4.2 experiment, we utilized 7 male celebrities in the text prompt, 6 reference clothing items, and 9 control strengths to generate 10 samples per control strength, resulting in a total of 7560 images. The quantitative result corresponds to Figure 11 and is listed in Table 1. We will elaborate on the quantitative findings in conjunction with the qualitative results.

A.1 Control Strength Analysis

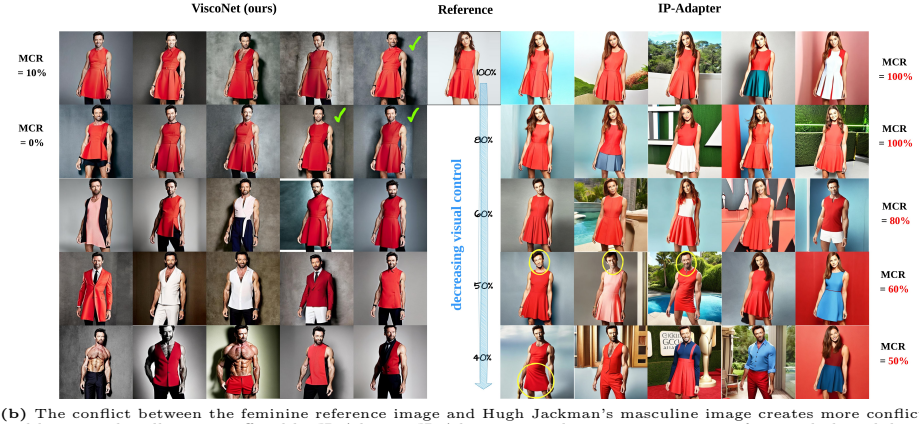
Strength	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
	CLIP score								
IP-Adapter	0.3066	0.3052	0.3003	0.2910	0.2729	0.2546	0.2231	0.1687	0.1606
Method 1 (ours)	0.3223	0.3230	0.3218	0.2993	0.2139	0.1846	0.1768	0.1670	0.1628
Method 2 (ours)	**0.3232**	**0.3254**	**0.3250**	**0.3218**	**0.3196**	**0.3064**	**0.2920**	**0.2544**	**0.2312**
	Mode Collapse Rate (MCR)								
IP-Adapter	0.05	0.06	0.11	0.19	0.32	0.45	0.67	0.99	1.00
Method 1 (ours)	**0.00**	**0.00**	**0.00**	0.12	0.74	0.97	1.00	1.00	1.00
Method 2 (ours)	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.08**	**0.15**	**0.40**	**0.57**
	MS-SSIM								
IP-Adapter	0.1248	0.1420	0.1705	0.2033	0.2452	0.2888	0.3700	0.4447	0.5107
Method 1 (ours)	**0.1514**	**0.1795**	**0.2312**	**0.3781**	**0.5043**	**0.5335**	**0.5498**	**0.5536**	**0.5409**
Method 2 (ours)	0.1507	0.1627	0.1886	0.2538	0.3661	0.4387	0.4742	0.5173	0.5183

Table 1: Reduced control strength results in higher CLIP scores and accuracy, translating to less mode collapse.

In Figure 15, we show experiment samples of text prompt *Hugh Jackman* and reference image 2, where we randomly sample 50% of the samples for various



(a) IP-Adapter suffer 100% mode collapse at control strength over 80% and unable to generate the target person *Hugh Jackman*. Its visual conditioning power is much weaker when it finally escapes mode collapse at lower control strength, and unable to generate the short pant, and correct clothing style and color. In contrast, we are robust against mode collapse and avoid much of the problems above suffered by IP-Adapter, and able to generate desired results ✓ at 100% control strength, preserving faithfulness of both the person identity and clothing appearance.



(b) The conflict between the feminine reference image and Hugh Jackman's masculine image creates more conflict and hence mode collapse as suffered by IP-Adatper. IP-Adatper struggles to generate correct faces and pleated dress patterns (circled in yellow) at weaker control strength. This does not affect our method.

Fig. 15: Comparing the effect of control strength on re-identification task. IP-Adapter suffers much more severe mode collapse and struggles to create perfect image balancing the reference image and text prompt of *Hugh Jackman*.

control strengths from both our and IP-Adapter. With 100% strength, although IP-Adapter can reconstruct the reference image, it suffers 100%

Overall, our method is effectively mode collapse free at 60% while IP-Adapter still 1 has 67% MCR. As shown in Figure 15, although high control strength introduces some mode collapse to our method. However, we can still generate high-quality images, preserving visual conditioning and a person's identity.

A.2 Further Quantitative Comparison

We further explore qualitative results in this section. Unlike Figure 7-10, where we slide along the control strength on the same random seed to demonstrate latent space discontinuity, we extend Section A to present the best samples across all control strengths from both methods for direct comparison, as shown in Figure 16-18.

(a) Visual reference taken from the unseen test dataset.

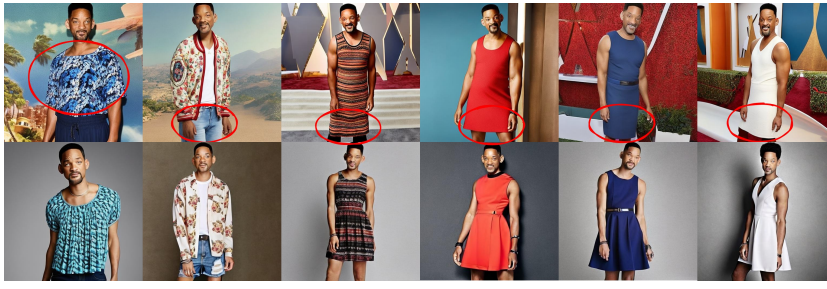
(b) Unlike other movie stars with more diverse costumes, Prince Charles' limited clothing range presents the toughest challenge. (Top) IP-Adatper cannot produce any image of Prince Charles wearing the reference clothing. (Bottom) despite the extreme data gap, our method can produce reasonable images.

Fig. 16: Most challenging example in re-identification task - Prince Charles.

Among all the celebrities mentioned in the text prompt, *Prince Charles*[1] - known for having a limited wardrobe of formal attires in public images - presents the greatest challenge to the generalization capability of the models. IP-Adapter encounters difficulties and fails to generate any image of Prince Charles in casual or feminine clothing, as depicted in the reference image (Figure 16). In contrast, our method achieves reasonable success despite the monumental challenge. Figure 17 - 18 shows samples from the rest of the text prompts used in the experiment. Overall, IP-Adapter needs to have much-lowered control strength to escape mode collapse, resulting in loss of fidelity in clothing to the reference images, including the incorrect length of pants or dress, wrong color and pattern, i.e., loss of the pleated dress pattern, it previously able to generate (Figure 15b).

[1] Stable Diffusion was trained on dated data before Prince Charles ascended to be king, so we adhere to his old title in the experiment.

(a) Visual reference taken from the unseen test dataset.



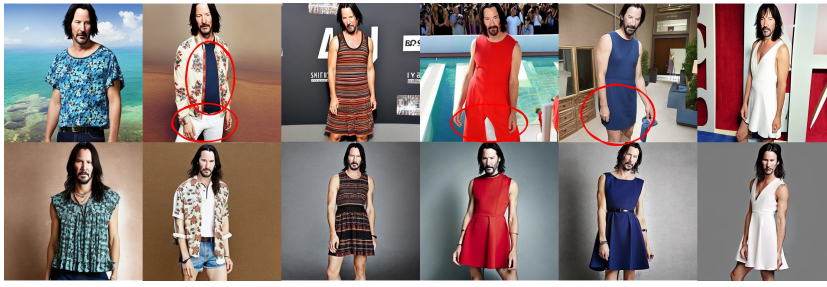
(b) Will Smith: (top) IP-Adaptor showing incorrect clothing color, length, or style (no pleated dress pattern). (bottom) Ours

(c) Dwayne Johnson: (top) IP-Adaptor (bottom) Ours

(d) Hugh Jackman: (top) IP-Adaptor (bottom) Ours.

Fig. 17: Re-identification comparison with IP-Adapter.

(a) Visual reference taken from the unseen test dataset.



(b) Keanu Reeves: (top) IP-Adaptor (bottom) Ours.

(c) Robert Downey Jr.: (top) IP-Adaptor (bottom) Ours

(d) Tom Cruise: (top) IP-Adaptor (bottom) Ours

Fig. 18: Re-identification comparison with IP-Adapter.

Fig. 19: Putting our images together shows the consistency of our method in delivering celebrity re-identification.

B Versatile Human Image Generation Task

B.1 Re-identification (visual prompt)

Figure 20 shows by conditioning on face and hair images, our method generates realistic people with diverse skin tones and body shapes correctly matching the faces despite the DeepFashion dataset consisting of more than 90% of female images, predominately fair-skinned women.

Fig. 20: Re-identification with a visual prompt.

B.2 Stylization

Figure 21 and Figure 22 show that our visual conditioning is effective across many image domains in creating a desired person's appearance, including various painting styles and also 3D objects such as statues, sculptures, toys, and 3D graphics. Some image domains have distinctive characteristics with considerable divergence from real photos, such as cartoons with disproportionate bigger heads, which can lead to a higher mode collapse rate. We circumvent this by removing the face mask to create results such as in Figure 21l and Figure 22l.

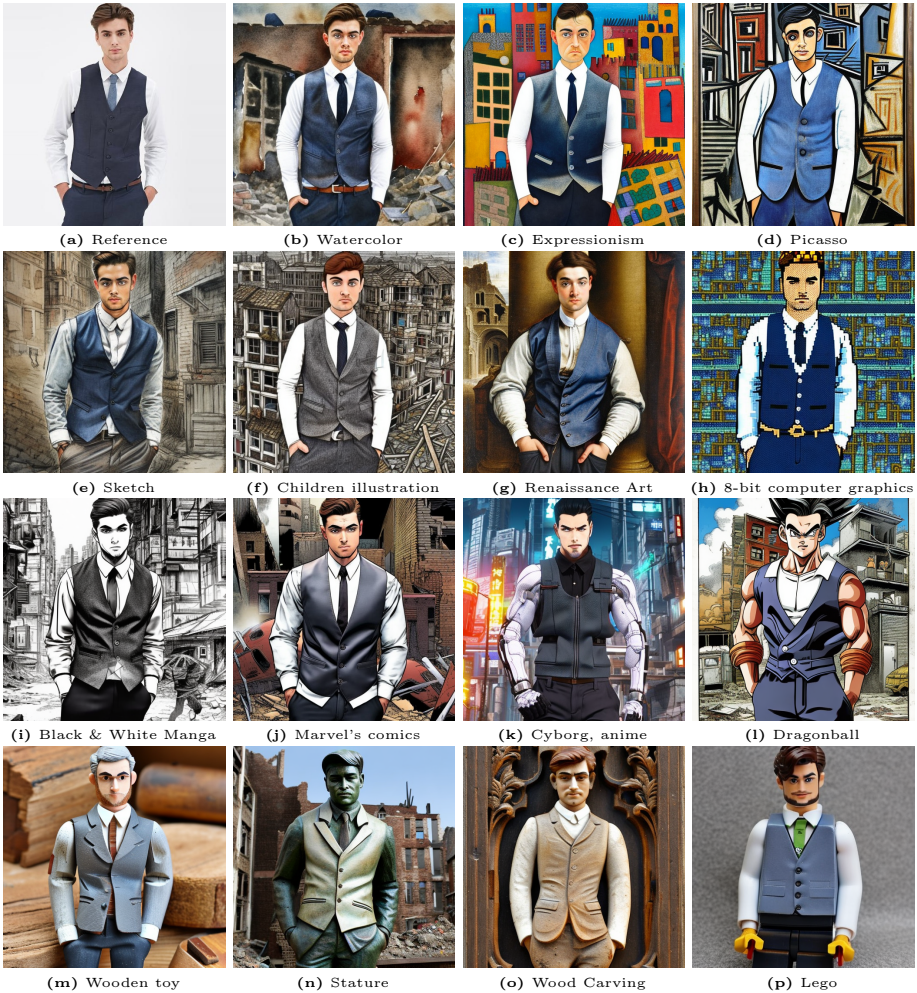
(a) Reference (b) Cartoon (c) Color sketch (d) Van Gogh

(e) Pencil sketch (f) Portrait de Messieurs (g) Cubism Art (h) Picasso

(i) Chinese ink painting (j) Japanese paper art (k) cyberpunk anime (l) Disney's Frozen

(m) Minecraft (n) Shaun The Sheep (o) Barbie doll (p) Lego

Fig. 21: Stylization. Text prompt: *"a woman, in farm."*



(a) Reference (b) Watercolor (c) Expressionism (d) Picasso

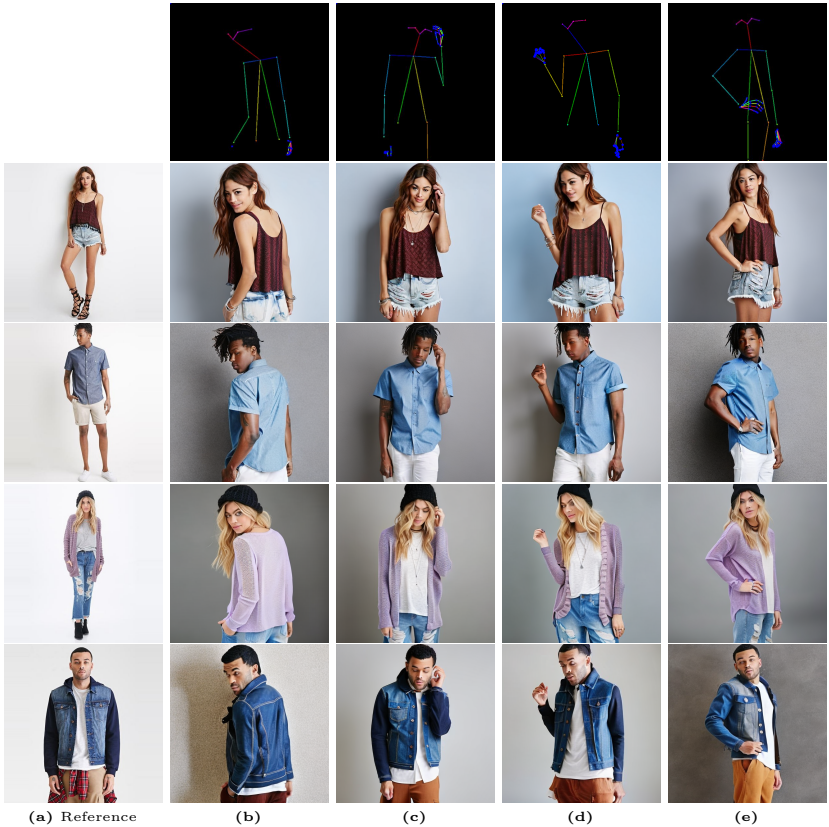
(e) Sketch (f) Children illustration (g) Renaissance Art (h) 8-bit computer graphics

(i) Black & White Manga (j) Marvel's comics (k) Cyborg, anime (l) Dragonball

(m) Wooden toy (n) Stature (o) Wood Carving (p) Lego

Fig. 22: Stylization. Text prompt: *"a man, in a derelict city."*

B.3 Pose Re-target



(a) Reference (b) (c) (d) (e)

Fig. 23: Pose Transfer from (a) reference person to new poses in (b)-(e)

B.4 Virtual Try-on

Figure 24 demonstrates how we perform fashion virtual try-on using visual and text prompts. Figure 25 illustrates the culmination of our methods, showcasing the seamless integration of re-identification, virtual try-on, and pose re-target.

(a) (b) (c) (d) (e) *"yellow jacket"*

Fig. 24: High-resolution virtual try-on with real-world background. (Top) reference fashion for visual conditioning. (Bottom): virtual try-on results.

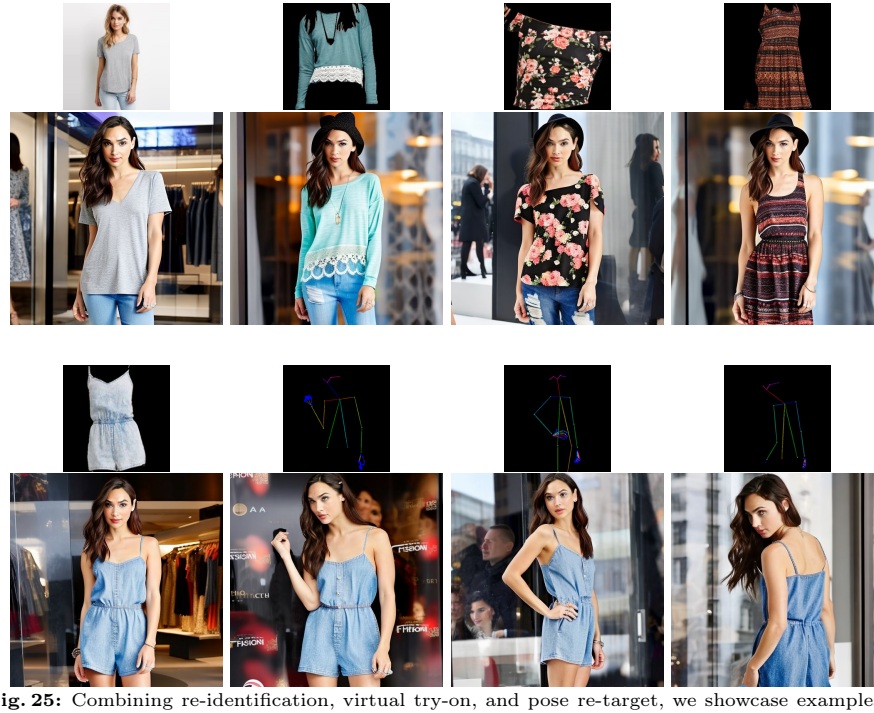


Fig. 25: Combining re-identification, virtual try-on, and pose re-target, we showcase examples of posing fashion with celebrity avatars.

C Quantitative Result

C.1 Section 4.1: Mode Collapse Quantitative Result

Table 2 shows the quantitative results corresponding to Figure 6 in Section 4.1 - Mode Collapse and Control Strength. Our method produces a higher CLIP score than the baseline at various control strengths, indicating less mode collapse. This is more evident in CLIP accuracy; at control strength 0.5, we achieve 100% (or 0% MCR) while baselines have only 46% and 63% ControlNet and IP-Adapter, respectively.

Strength	0.0	0.2	0.3	0.4	0.5	0.6	0.8	1.0
				CLIP score				
ControlNet	0.2720	0.2620	0.2440	0.2340	0.2300	0.2300	**0.2240**	**0.2260**
IP-Adapter	**0.2900**	0.2920	0.2780	0.2620	0.2360	0.2120	0.1780	0.1900
ViscoNet(Ours)	0.2860	**0.2940**	**0.2920**	**0.2900**	**0.2800**	**0.2660**	**0.2420**	0.2220
				CLIP accuracy				
ControlNet	0.8660	0.7720	0.7000	0.6180	0.4620	0.5500	0.5020	**0.5760**
IP-Adapter	0.9800	0.9700	0.8800	0.7500	0.6300	0.4100	0.1500	0.2100
ViscoNet(Ours)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.9000**	**0.7000**	**0.5760**
				Pose accuracy (OKS)				
ControlNet	0.0880	0.4139	**0.6610**	**0.8305**	**0.8596**	**0.8852**	**0.9223**	**0.9348**
IP-Adapter	**0.5379**	**0.5412**	0.6060	0.6813	0.7546	0.8074	0.9010	0.9298
ViscoNet(Ours)	0.0446	0.1654	0.3869	0.6580	0.7845	0.8253	0.8824	0.9102

Table 2: Reduced control strength results in higher CLIP scores and accuracy, translating to less mode collapse.

Figure 26 shows the breakdown of CLIP accuracy across the image styles in Table 2. Based on the same Stable Diffusion model, all models have shown the highest mode collapse rate in Van Gogh's painting style, while Ukiyoe is the least affected.

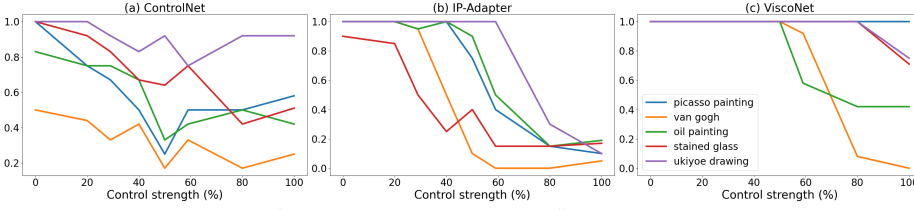


Fig. 26: CLIP accuracy - comparing different image styles.

C.2 Section 4.3: Human Evaluation Result

We further conducted a more extensive scale user study on Amazon Mechanical Turk (AMT) to measure the real-life preferences between our model and the HIG baseline approaches. We perform a 4-way comparison, asking workers to select their best preference from randomly shuffled samples, as shown in Figure 27.

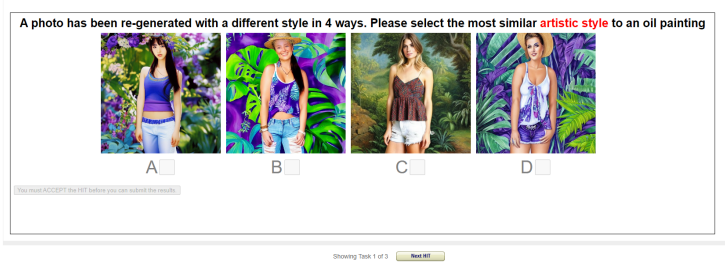


Fig. 27: Screenshot of user study presented to users for evaluating the quality of the stylization against the three baselines.

	Human Evaluation				
Image Styles	HumanSD	ControlNet	T2I-Adapter	**ViscoNet (Ours)**	Ours (%)
Ukiyoe	27	32	4	**37**	**37%**
Cyberpunk anime	23	13	21	**41**	**41%**
Stained glass	0	32	23	**45**	**45%**
Van Gogh	2	13	9	**76**	**76%**
Picasso	0	13	42	**45**	**45%**
Oil Painting	9	11	7	**73**	**73%**
Disney	5	23	5	**67**	**67%**
Total	77	139	111	**384**	
Average	9.43%	19.9%	15.9%	**54.9%**	

Table 3: Our method scores the highest in human evaluation, proving its ability to generate good-quality, diverse image styles.