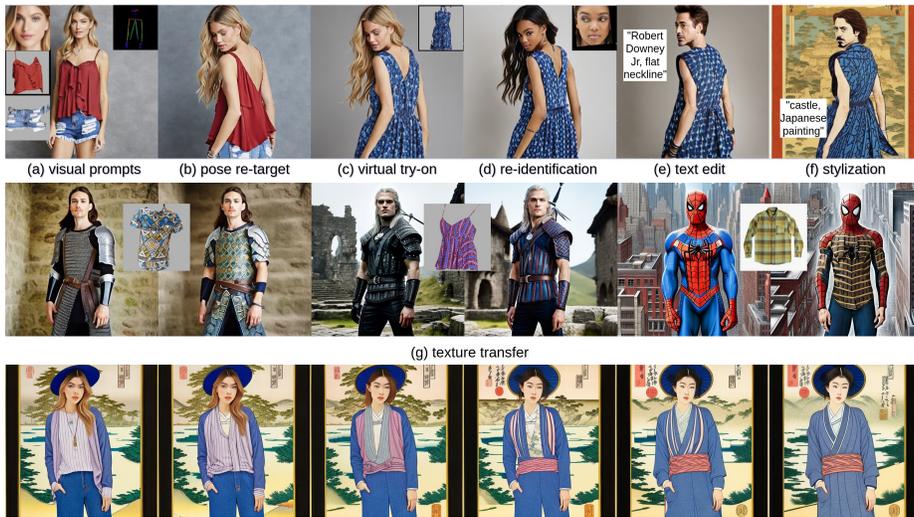


ViscoNet: Bridging and Harmonizing Visual and Textual Conditioning for ControlNet

Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert

s.cheong, armin.mustafa, a.gilbert@surrey.ac.uk
University of Surrey, UK

Abstract. This paper introduces ViscoNet, a novel one-branch-adapter architecture for concurrent spatial and visual conditioning. Our lightweight model requires trainable parameters and dataset size multiple orders of magnitude smaller than the current state-of-the-art IP-Adapter. However, our method successfully preserves the generative power of the frozen text-to-image (T2I) backbone. Notably, it excels in addressing mode collapse, a pervasive issue previously overlooked. Our novel architecture demonstrates outstanding capabilities in achieving a harmonious visual-text balance, unlocking unparalleled versatility in various human image generation tasks, including pose re-targeting, virtual try-on, stylization, person re-identification, and textile transfer. Demo and code are available from project page <https://soon-yau.github.io/visconet/>.



(h) stylization and latent space interpolation: person appearance and clothing morph from (left) visual prompt to (right) text prompt "Japanese painting style"

Fig. 1: Our proposed **Visconet** demonstrates broad versatility in multimodal human image tasks including visual prompts, pose re-target, virtual try-on, re-identification using either text or visual prompt, texture transfer, stylization and latent space interpolation to perform human morphing.

1 Introduction

Diffusion models [14, 26, 29, 35] are powerful tools for generating realistic and diverse images and videos from various inputs. Among them, latent diffusion models (LDM) [32], more notably Stable Diffusion (SD) [32], have shown impressive results in text-to-image (T2I) synthesis, thanks to their high quality and open-source availability. However, relying solely on text as the input condition introduces several limitations, notably the challenge of providing a comprehensive description of an image. Furthermore, concept bleeding is a prevalent issue in T2I, as highlighted by works such as [4, 21], where the text becomes erroneously associated with incorrect subjects in the generated images. In human image generation (HIG), this misassociation may manifest in inaccuracies such as assigning the wrong clothing color or experiencing color spillover between clothing and the background, and vice versa.



Fig. 2: To motivate our work, this figure illustrates how increasing text complexity in ControlNet [46] can expose (c) domain gap and eventually lead to mode collapse in (d). IP-Adapter [43] also exhibits (e) catastrophic forgetting, resulting in the inability to generate a rich background. Both show the concept of bleeding by assigning the wrong color to clothing garments.

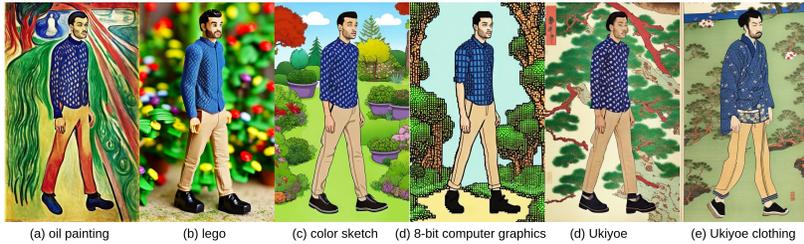


Fig. 3: Our method retains generative power of the T2I backbone in (a)-(d) various image styles and rich backgrounds while maintaining the person and clothing appearance, assigning correct clothing colors. In (e), we can control the level of stylization to expand it to the clothing styles.

While incorporating new conditioning factors, such as pose information, into a model for novel tasks necessitates retraining from scratch [8, 20, 23, 45], demanding substantial datasets and computational resources. Recognizing this challenge, recent methodologies like T2I-Adapter [25] and ControlNet [46] have introduced a pragmatic approach. They integrate a lightweight adapter branch [15] to encode spatial conditioning information, such as pose or segmentation maps, onto a frozen pre-trained T2I LDM backbone. In a more recent development, methods such as IP-Adapter [43] and MasaCtrl [2] extend this concept by introducing visual conditioning capabilities. However, they cannot control

pose independently and require a separate spatial adapter, introducing additional computational complexity to the overall architecture. However, training adapters on smaller and disparate datasets may introduce a domain gap with the frozen LDM model. As noted by [2, 20], this conflict between branches can manifest in the model’s inability to generate people following specified text and pose conditions. The situation may be exacerbated when multiple adapter branches are employed. Our work addresses this issue by striving to develop a lightweight adapter that accommodates both pose and visual conditioning. This singular adapter aims to excel in a spectrum of human image generation tasks, unifying functionalities currently achievable through utilizing distinct models.

We illustrate the domain gap and conflict of adapters in Figure 2 where ControlNet attempts to reconstruct reference images with increasing text complexity from (b) to (d). Figure 2c shows a sign of domain gap as dark-skinned people were not typical in ancient Japanese drawing (Ukiyoe style). We continue adding “khaki”, a more modern term, into the text prompt. The complexity eventually exposes the domain gap between ControlNet and T2I. As a result, ControlNet resorts to generating realistic people that it learned from its small training data (Figure 2d). This is a phenomenon we call **mode collapse (MC)**. Mode collapse has existed since GANs [13] but has not been discussed recently despite widely affecting recent diffusion model-based adapters. We are the first to study mode collapse in an adapter-based diffusion model systematically. There is currently no effective mechanism to control and manage this conflict; only when one of the conflicting texts, i.e., khaki or Ukiyoe style, is removed will it escape the stuck mode. Unfortunately, this restricts the image content that can be generated. The general solution is to train a more extensive dataset to close the domain gap. HumanSD [20] compiled a 1M image dataset, up from ControlNet’s 200k, while IP-Adapter uses 10M [43] and more recent Hyperhuman [23] ballooned to 340M! This is an inefficient use of computing resources, and as we will show, this is insufficient to eradicate mode collapse completely. Conversely, training on a limited dataset may lead to overfitting and, consequently, catastrophic forgetting. This is evident in the model’s diminished ability to generate diverse individuals, varied image backgrounds, or encompassing artistic styles as depicted by the input prompts. In contrast, our method trains only on about 50K images, many orders of magnitudes smaller than reference methods.

In this paper, we propose a novel architecture extending ControlNet [46], which we call **ViscoNet (Visual ControlNet)**, bridging and harmonizing visual and text conditioning. Our method’s ability to fuse and control the balance of both text and visual conditioning unlocks unparalleled versatility in HIG, which includes pose re-target (transfer), virtual try-on, person re-identification (face swap) with both text and visual, image stylization, textile transfer, and visual-text latent space interpolation to achieve morphing as shown in Figure 1. The summary of our contributions:

1. A lightweight one-branch adapter architecture for spatial and visual conditioning.
2. Excellent ability to control and harmonize text and visual prompts, significantly mitigating mode collapse and empowering various HIG capabilities.

3. Our training with feature masking effectively preserves the backbone model’s generative capabilities on a small dataset, mitigating catastrophic forgetting.

2 Related Works

Human Image Generation in early days uses GANs [24, 31, 44, 47, 49] predominately taking pose and reference image as input conditions to perform pose re-target and virtual try-on. Later, architectures based on transformer [38] e.g., [10, 30] and notably diffusion models [1, 14, 20, 26, 29, 32, 35] increasingly became the mainstream image generation methods. However, they used only either text or image but not both as input modality, limiting the controllability. Therefore, text prompt is added to specialist HIG models [7, 8, 19] to enrich the finer level of control. These models are typically trained from scratch on small datasets resulting in overfitting and an inability to generalize to generate realistic images in diverse, real-world scenarios.

Visual Conditioning. Image personalization methods [11, 34] explore finetuning text vocabularies to define specific identities. [5, 12] follow the same idea, while [6, 18, 36] leverage large-scale upstream training to eliminate the need for test-time finetuning. These methods use text to control visual aspects rather than images as input conditioning. In HIG, UPGPT [8] pioneered visual conditioning in the T2I diffusion model by concatenating visual tokens alongside text tokens and pose tokens. However, it changes the model architecture and unable to re-use the pre-trained model weights.

Adapter. More recently, adapter modules and lightweight models have been added to pre-trained, frozen diffusion models for faster finetuning requiring less data; among them are ControlNet [46], T2I-Adapter [25]. However, as they add the learned feature spatially to the UNet’s multi-resolution layers in the diffusion model, the control signals are limited to the spatial dimension. Although the T2I-Adapter demonstrates the use of reference images for visual conditioning, it is constrained to the overall artistic style of the image. MasaCtrl [2] is a tuning-free method that injects masked self-attention features from a reference image in the T2I denoising step. IP-Adapter [43] uses a separate cross-attention map for image conditioning to be added to the textual attention map. The balance can be adjusted using a weighted average between the two attention maps. Both IP-Adapter and MasaCtrl are conditioned on a single image for a global image, lacking fine-grained visual conditioning. Uni-ControlNet [48] supports both global and local image but still employs dual-branch design. InstantID [39] is based on IP-Adapter’s architecture, with the main difference being swapping the CLIP image encoder with a specialized face encoder. While they focus on human face, our method exhibits a broader capacity, generating full human body with higher complexity.

Dancing Avatar. This group of models re-purposes T2I into image-only-conditioning to reconstruct humans for dancing avatar videos faithfully. They sacrifice the T2I’s text capability and are not directly comparable to our method. Nevertheless, we scrutinize their pose-and-visual methods. Disco [40] uses ControlNet to inject static image background signal. To ensure visual consistency

of the moving foreground person, it applies a visual signal to cross-attention of UNet in image-to-image SD variant [27], which requires re-training. MagicAnimate [42] and AnimateAnyone [16] use a dedicated adapter branch to encode visual information to be fused with UNet using cross-attention.

Overall, existing methods [2,16,25,40,42,43,46,48] employs multiple adapters for simultaneous pose (e.g. ControlNet) and visual control (e.g. IP-Adapter). Our method introduces improvements over a single ControlNet to offer both pose and visual control, saving computational requirements and potentially mitigating conflicts introduced by multiple branches.

3 Method

3.1 Preliminaries

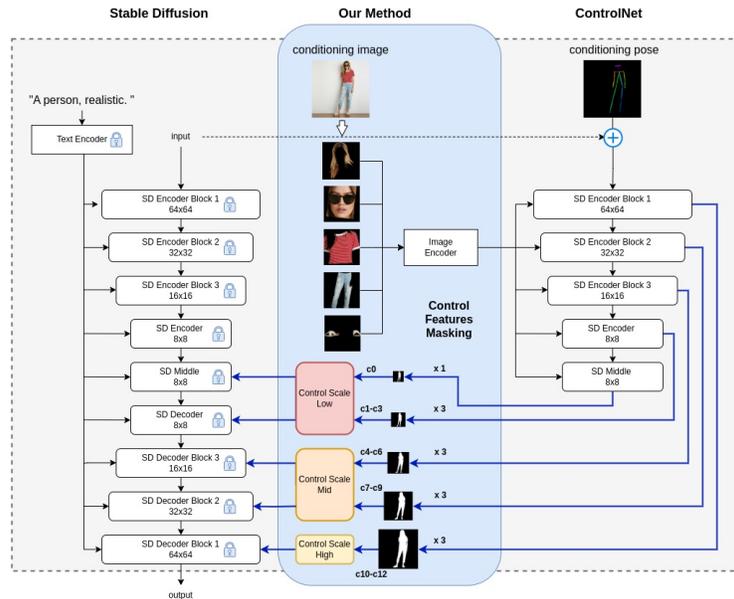


Fig. 4: Architectural diagram showing our contribution concerning backbone LDM and ControlNet layers. We omit time embedding, zero convolution, and some blocks from the ControlNet diagram [46] for simplicity.

Stable Diffusion (SD), a backbone LDM [32], and a ControlNet model [46] are shown in the left and right block in Figure 4. SD uses a UNet [33] as the denoising network and progressively refines the input noise into latent variables that can be reconstructed into realistic synthetic images, relying on understanding intricate image distributions. The words within a text prompt are decomposed into smaller subword units and tokenized and encoded with a CLIP [28] text transformer [38]. The text embedding is injected into the cross-attention layers in UNet, serving as the sole conditioning in image generation. The loss function of the LDM is:

$$\mathcal{L}_{MSE} := \mathbb{E}_{z,c,t,\epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_{\theta}(z_t, t, c)\|_2^2] \quad (1)$$

where c is the text conditioning token, t is the diffusion time step, and z is the latent variable (denoted as input in Figure 4).

Instead of training from scratch, ControlNet [46] adds a learnable branch parallel with a now frozen pre-trained LDM, as shown on the block on the right in Figure 4. The branch consists of an identical LDM UNet encoder copy, sharing the same latent noise input and text embedding. It learns to control pose conditions by adding skeleton image features into the latent noise input at the branch input. ControlNet generates spatial control signals and adds them to the SD decoder across multiple spatial resolutions.

3.2 Replace Text with Visual Prompt

That ControlNet’s sharing of the exact text embedding with the LDM is unnecessary when learning the spatial condition. Their mandatory use of text-image pairs in training places an excessive burden on data collection and annotation to the specific image and text styles. The text entanglement also increases potential conflict between the branch and LDM [20]. Therefore, in our architecture, we remove the text prompt from ControlNet to sever the entanglement and replace it with a visual prompt. Unlike [40, 43] that use a single reference image for overall visual conditioning, we use multiple images consisting of segmented body parts (e.g. hair, face, top clothing, bottom clothing) for fine-grained visual control on individual clothing garment pieces (Figure 1a-e).

The de-facto image encoding method for the diffusion model uses a CLIP image encoder to extract a global image embedding, but this is insufficient in capturing intricate image details. Therefore, we utilize the larger dimension local CLIP embedding before the pooling layer. We project the local CLIP embedding of individual images using a linear layer into length N and concatenate them like the text tokens they replace. N can be adjusted based on the number of conditioning images and the text token length limit of ControlNet, and we use $N=8$ in this work. The linear layer consists of only 2K parameters. It is the only additional trainable parameter introduced by our method, 10,000 times fewer than IP-Adapter’s 22M parameters. Controlling all the human information (pose and visual appearance) within the single adapter branch creates effective disentanglement from the LDM to avoid conflict. For example, “ripped jeans” may conflict with “Picasso painting”; having both in a text prompt could trigger mode collapse in ControlNet. Instead, we can avoid this by removing “ripped jeans” from the text prompt and replacing it with an image in the visual prompts.

3.3 Control Feature Masking

DeepFashion [51] is a popular and de-facto dataset in many HIG tasks in machine vision literature. However, all the images consist of plain studio backgrounds, which will overfit and severely restrain LDM generative capability, resulting in dull and bland image backgrounds. This phenomenon is observed with IP-Adapter [43] in Figure 2e despite it being trained on a large dataset of 10M images. To tackle this issue, we apply a binary human silhouette mask to the control signals originating from our adapter branch before injecting it into the

LDM. This eliminates unwanted image backgrounds leaking into and causing overfitting in the LDM. The disentanglement between foreground people and background contributes to reducing the image domain gap with the LDM. This improvement empowers our method to harness LDM text capability to generate vibrant backgrounds in various artistic image styles despite training only on a relatively tiny dataset (only 52K) of images with plain backgrounds. In Section 5, we delve into further analysis, demonstrating that feature masking is essential during training rather than used solely during image sampling.

The feature mask is also applied to the LDM loss function (Equation 1) at the *output* in Figure 4. The training loss backpropagates via the frozen LDM to train the model. This approach is akin to [8, 20], although they use it to assign weight loss to different body segmentation parts rather than masking a region entirely. We add masking to the LDM loss function 1:

$$\mathcal{L}_{MSE} := \mathbb{E}_{z,c,v,t,\epsilon \sim \mathcal{N}(0,1)} [\|\mathcal{M} \odot (\epsilon - \epsilon_\beta(z_t, t, c, v))\|_2^2] \quad (2)$$

where ϵ_β is the model, v is the image embedding, \odot is the element-wise multiplication, and $\mathcal{M} \in \mathbb{R}^{H,W}$ is the binary mask resized to resolution (H, W) of the LDM output. Although text conditions are not used in the model, they are used by LDM in training and, thus, are included in the equation.

3.4 Harmonizing Text and Visual Influence

The versatility of our approach in performing diverse human image generation tasks arises from its ability to seamlessly integrate and regulate the balance between visual and textual conditioning. We achieve this by multiplying scalar values [0.0, 1.0] with the control features. Scaling control signals is commonplace in adapter-based approaches, but our novel model architecture unlocks unprecedented effects not observed in existing methods. Despite innovations adopted to reduce data conflict between the adapter and the LDM, mode collapse can still happen in challenging image styles. In this scenario, we can decrease the control signal strength to weaken the visual prompt strength to escape the mode collapse. As we will show, the application of this approach has no discernible impact on mitigating mode collapse in ControlNet [46] and other spatial conditioned models [20, 25]. This is attributed to the fact that its control signals exclusively influence pose conditioning, whereas the root causes of the conflict lie in the image domain gap and text entanglement. In contrast, our innovative architecture, which involves the separation and subsequent bridging of text and visual conditioning, empowers us to dynamically adjust their balance, thereby enabling latent space interpolation (Figure 1h) and eliminating mode collapse.

On the other hand, IP-Adapter [43] supports visual prompts, and it can adjust the text-visual balance by changing the scales of the respective cross-attention map, but the effect is global to the image. For example, tipping the balance away from the visual prompt of a realistic photo of a man towards the text prompt “a girl, Chinese ink painting” would result in the global transformation of a modern man towards a Chinese girl wearing period Chinese clothing in Chinese ink painting style. Our method can apply different scaling at each

multi-spatial resolution to customize at different image levels. This is demonstrated in Figure 1f, which depicts only the image’s artistic style while retaining the person’s identity and appearance, and Figure 1h, which shows the morphing only of the person, leaving the background essentially unchanged.

For the sake of discussion, the 13 individual control strength scales (c0-c12) can be roughly grouped into three blocks - Low Blocks (LB), Mid Blocks (MB), and High Blocks (HB) arranged hierarchically from low to high spatial resolution. We can adjust their values separately to create different effects. Through experimentation, we observed that LB exerts negligible influence and can effectively be set to 0. The MB is the most influential in overall visual appearance styles among them. HB regulates fine image texture, aligning with our expectations for image hierarchy control. Setting HB alone yields the notable outcome of transferring only the texture of the visual prompt (Figure 1g). We can also constrain our control to pose only by setting c4 to 0.5 while leaving others 0.0, allowing using text prompts to control the whole person’s appearance.

3.5 Training Setup

To train the model, we employ 52K-images DeepFashion In-shop Clothes Retrieval dataset [51] and adopt the train-test split proposed by [50] for the pose transfer task, padding the images to the size of 512×512 . Pose information is extracted using OpenPose [3] to create body-and-hand skeleton images, and we use pre-segmented fashion images from [8]. We employ a simple text prompt of “a person” for all the images. This serves two purposes: first, the neutral description avoids potential conflicts with the LDM, proving our method does not need to annotate text to match the style of the LDM carefully. Secondly, it acts as an unconditional text embedding, enabling users to amplify the desired visual effect using positive prompts, negative prompts, and guidance scales [9].

Many adapter models are based on pre-trained SD or similar-sized models. Thus, we also performed our experiments using SD2.1 [37] for a fair comparison. We initialize our adapter branch by copying frozen weights from the SD. However, since the cross-attention input has shifted from global CLIP text embedding to local CLIP image embedding, we re-initialize the weights in the cross-attention layer at the start of training. All weights in the SD, CLIP text, and image encoders are frozen. We use CLIP image encoder *clip-vit-large-patch14* [17]. We trained the model on a single desktop GPU GTX 3090 for 2 epochs, using a batch size of 4 with four gradient accumulations per batch, resulting in an effective batch size of 16. We retained the remaining configurations from [46].

3.6 Image Resolution

We use an image resolution of 512×512 throughout the paper. In our experiment, we utilized $3/4$ length to full-body images, resulting in smaller human faces within the images. The stringent demand for high pixel density per latent variable can lead to suboptimal face construction [8] compared to high resolution face images generated by [39, 43]. This inherent limitation is a characteristic drawback of the LDM rather than a weakness of our method.

4 Experiments

In Section 4.1, we perform an in-depth study of the effect of control strength on mode collapse as observed in image artistic styles. We show that visual prompt methods (IP-Adapter and ours) effectively reduce mode collapse compared to ControlNet. Then, in Section 4.2, we perform further, more challenging experiments in person re-identification to show our method has superior text-visual harmonization capability compared to IP-Adapter. Lastly, we performed large-scale human evaluation in Section 4.3 to prove our image quality over SOTA HIG models.

4.1 Mode Collapse and Control Strength

In this section, we examine the prevalence of mode collapse and its impact on existing spatial and visual adapter models compared to our proposed model. In Figure 5, conditioned on the same human pose, we generate images of *Picasso style* at different control strengths. ControlNet does not have visual input. Therefore, we use text to describe the person’s appearance and background, which include conflicting word “*ripped jeans*” to invoke mode collapse. In this example, mode collapse happens to both ControlNet [46], IP-Adapter [43], and our proposed ViscoNet at a control strength of 80%, as observed with the realistic person and background in Figure 5e. However, our method has quickly escaped mode collapse at a control strength of 60% (Figure 5d), as at this point, the visual conditioning is still effective, maintaining the overall clothing styles and colors. We can also observe the harmonized transition towards the desired image style as reduced control strength tips the balance towards text prompt depicting “*Picasso*”(Figure 5a). Both reference methods only managed to escape mode collapse at around 40% (Figure 5b), which has considerably weakened pose or visual control for ControlNet and IP-Adapter, respectively.



Fig. 5: Effect of control strength (%). Compared to ControlNet and IP Adapter, our method can escape mode collapse faster, generating a harmonious image style while maintaining good visual control.

We confirm our qualitative observation with quantitative results. Like [43], we measure the effectiveness in generating the correct image styles by employing the CLIP similarity score between the text prompt and the generated image. A high CLIP score indicates a low or absence of mode collapse. We measure control effectiveness by measuring pose accuracy using the Object Keypoint Similarity (OKS) standard in MSCOCO challenge [22]. We will also introduce new metrics to measure and interpret mode collapse better. In this experiment, we selected 5 image styles - Picasso, Van Gogh, oil painting, Ukiyoe, and stained glass that are more likely to conflict with modern clothing. We generated 20 samples at each control strength (over 5000 images). The results are plotted in Figure 6, and we include the entire table in the appendix.

At 100% control strength, IP-Adapter lost most of its text capability, including changing image style (Figure 2e), resulting in the lowest CLIP scores (Figure 6a), indicating substantial mode collapse. The CLIP score for ControlNet remains constant in regions above 40%, whereas our method exhibits linear improvement in the same range. Both visual adapters effectively reduce mode collapse by using weaker control strength as indicated by weaker pose accuracy (Figure 6c). Our method consistently outperforms IP-Adapter in CLIP score at every control strength. It is worth noting that the IP-Adapter maintains its pose control for control strength <40% as they use separate adapters for pose control. Their drop of <40% is attributed to inaccuracy in pose detectors’ recognition of humans in artistic painting. Our quantitative results align with qualitative observations, establishing our method’s superior interpolation capability and ability to minimize mode collapse.

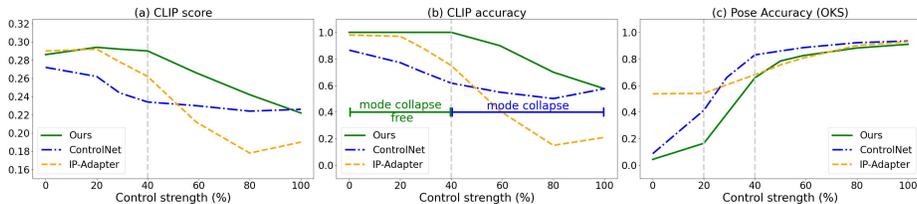


Fig. 6: (a) Reducing control strength alleviates mode collapse, our method can escape mode collapse faster, retaining better pose and visual control (b) CLIP accuracy provides better interpretability of mode collapse (c) Level of control as measured by pose accuracy. Visual prompting methods have slightly weaker pose control.

While CLIP scores are effective, their limitation lies in the lack of interpretability regarding the degree of mode collapse. Additionally, the absence of a standardized CLIP model within the machine learning community introduces variability, making cross-model comparisons challenging. Given these challenges, we explore alternative metrics for a more comprehensive evaluation. As mode collapse is an inherent discrete state, we employ CLIP binary classification ($CLIP_{acc}$) by comparing CLIP image embedding to two CLIP text classes - [image style], “real photo”. More generally, two modes are compared - target mode and stuck mode. In other words, we detect mode collapse if the image is classified as a real photo when it was supposed to be in the target image style. As shown in Figure 6b, $CLIP_{acc}$ correlates well to the CLIP similarity score but provides a normalized score easier for interpretation and enhanced robustness

against CLIP model variation. We define **mode collapse rate (MCR)** as :

$$MCR := 1 - CLIP_{acc} \quad (3)$$

Mode collapse is a phenomenon that occurs randomly, depending on the prompts and random seeds applied. Consequently, the MCR is a batch statistic that reflects the overall method performance. In Figure 6b, we achieve mode collapse free at a control strength of 40% (MCR=0% or $CLIP_{acc} = 100\%$) while IP-Adapter reaches that state later at much-weakened control strength.

4.2 Re-identification

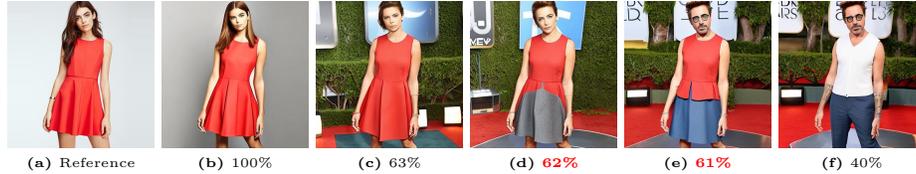


Fig. 7: IP-Adapter showing the transition from the reference image to text prompt “Robert Downey Jr.” by reducing control strength. There exists a big domain gap between (d) and (e).



Fig. 8: Method 1 - with head mask, smoother transition with smaller mode gap between (c) and (d).

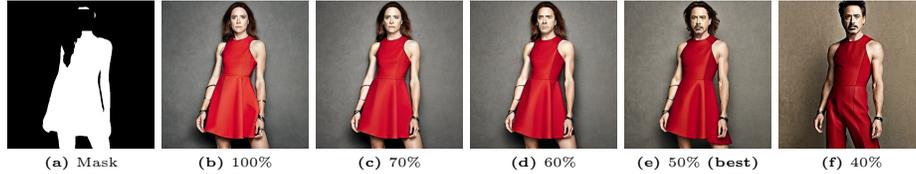


Fig. 9: Method 2 - without head mask. Smooth transition with (e) achieving good balance to deliver the desired result. The face and hair mask are detected and removed by segmentation tool. Our method has good tolerance over the mask region and does not require it to be pixel-accurate.

We formulated a demanding task to scrutinize an extreme instance of domain gap and assess the qualitative efficacy of visual prompting methods in addressing such challenges. In this task, the goal is to transform the person’s identity in the reference image into the person depicted in the text prompt, all while preserving the original clothing depicted in the reference image. In Figure 7, we show that decreasing control strength in IP-Adapter morphs the face towards the target (Robert Downey Jr.) at the expense of clothing faithfulness (red dress). A small control strength change between Figure 7d and 7e causes a significant shift in the image, indicating a big domain gap it fails to bridge harmoniously.

This common problem also affects our default configuration Method 1, which uses full human tasks. It achieves good results close to the target as shown in Figure 8d. Through extensive experimentation, we discovered that the face has

disproportionately influenced the entire image generation process. Consequently, it becomes imperative to substantially reduce the control strength (to around 40% in this example) to mitigate the impact of the face, albeit at the expense of visual control. Leveraging our novel architecture, we can effectively bridge this gap by selectively excluding the face from the feature mask, as shown in Figure 9 (Method 2). In essence, this action prevents the control signal from reaching the face region of the LDM. We tried applying a similar approach to the IP-Adapter by masking off the head from the reference image in pixel space, but this proved ineffective. This underscores the efficacy of our novel architecture in harmonizing text-visual controls. This has also proved useful in escaping mode collapse in certain challenging styles in stylization tasks (see appendix).

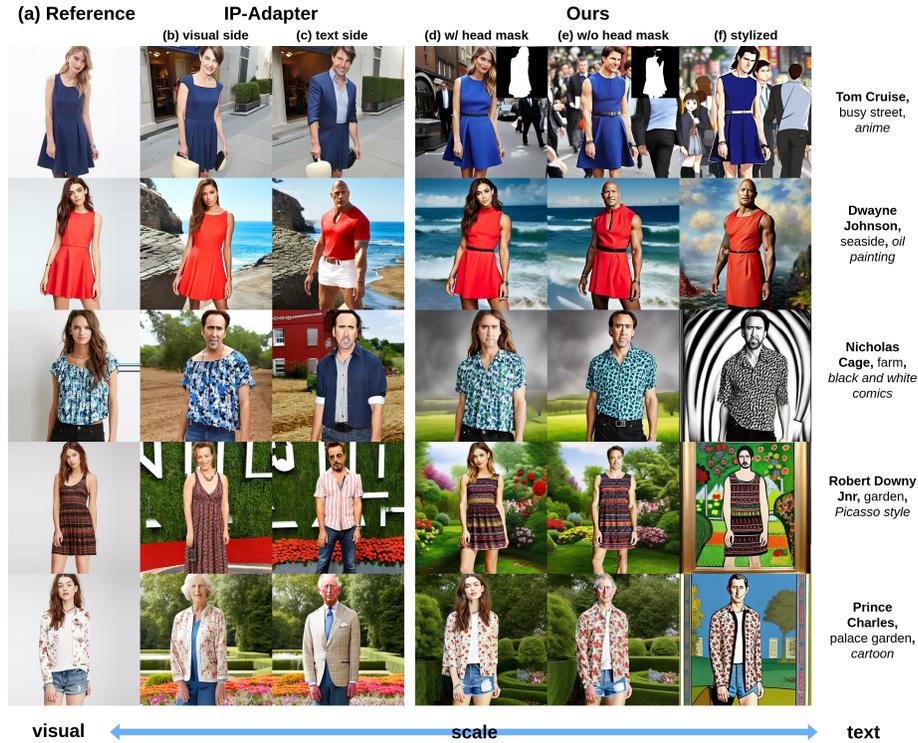


Fig. 10: Challenging re-identification task to transform female in (a) reference image to male celebrities depicted in text prompt. We included an additional stylization step (not included in the result) to demonstrate our ability to bridge the domain gap.

We generated over 5000 images from each method to perform the quantitative study; some test samples (input image and text) are shown in Figure 10. We include the image background and style to demonstrate our capability to maintain a constant background and bridging domain gaps across multiple dimensions to achieve stylization. We do not include them in our experiments as the objective is the foreground person identity and clothing. The experiment results are summarized in Figure 11 (full table in appendix). The presence of steep change in CLIP score (Figure 11a) and MCR (Figure 11b) with our Method 1 proves the evident domain gap within the 30%-50% control strength range. However,

removing the face from the mask in Method 2 drastically improves performance, outperforming IP-Adapter considerably. On the other hand, we measure effectiveness of visual control with image similarity score MS-SSIM [41] (Figure 11c). Method 1 (and 2) is consistently higher than IP-Adapter in MS-SSIM, suggesting more faithful visual appearance once escaping mode collapse (Figure 7 and 8).

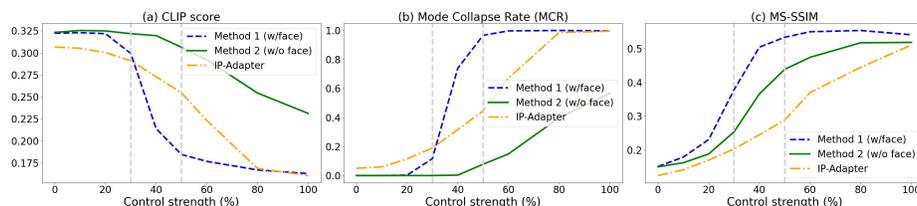


Fig. 11: Quantitative result showing the effectiveness of our method to escape mode collapse in the challenging re-identification task.

4.3 Generating Diverse Human Image Styles

We performed large-scale human evaluation comparing specialist SOTA pose-guided HIG models HumandSD [17], ControlNet [46], and T2I-Adapter [25]. In this experiment, we generate 1400 images evenly across seven image styles and the models (Figure 12). We use text prompts to describe clothing for reference methods and visual prompts for our process. In each test sample, 221 human evaluators were randomly shown a sample from each model and asked to pick one that best matches the text prompt. The majority, 55% of 700 responses (full table included in appendix), prefer our samples, proving overall superiority in image quality and visual control.

5 Ablations

Necessity of Feature Mask in Training. Catastrophic forgetting can be demonstrated using the DeepFashion dataset; in our initial experiments, we applied feature masking to the training loss function but excluded it from the control signals. However, applying the feature mask post-training is ineffective, as shown by the pale and dull background in Figure 13a -13c. In particular, the artifact (circled in red) in Figure 13c gives a clear indication of leakage of background originating from padding artifact uniquely caused by our dataset pre-processing error as shown in Figure 13d. Evidently, our method of applying feature masking in training produces a vibrant background (Figure 13e-13g), demonstrating that our method is effective in avoiding catastrophic forgetting.

CLIP Local Image Embedding Captures Fine Texture. We experimented with two image embedding methods for visual conditioning - global and local CLIP image embedding. Figure 14 shows that local CLIP embedding used in our method is better at capturing fine texture details.



Fig. 12: All reference methods have the purple clothing color spread into the forest background, while our method avoids this problem and can generate a vibrant and diverse background.

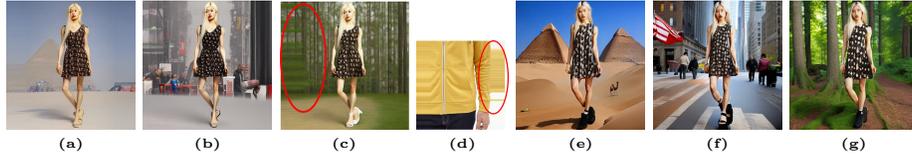


Fig. 13: Without feature masking in training : (a)-(c) pale, dull background (d) padding artifact from dataset. Using feature masking in training: (e)-(g) vibrant colors and rich background.



Fig. 14: Local CLIP image embedding used in our method can capture fine texture details. (left) local embedding (mid) visual prompt (right) global embedding.

6 Limitations

Like IP-Adapter, the clothing color in generated images is shaped by the inherent randomness in the initialized latent variables of the LDM backbone. While visual prompting proves effective with our method, attaining consistent and faithful image reconstruction necessitates careful selection of random seeds. On the other hand, by design choice, our model’s visual prompting method learns only the foreground people and leaves the background generation to the LDM backbone. Conversely, pose transfer requires perfect reconstruction of the image background solely from the reference image. Consequently, we refrained from conducting a large-scale evaluation in virtual try-on and pose transfer tasks. Nevertheless,

through careful random seed selection, we can still generate high-quality virtual try-on, pose transfer, and face swap images, as included in the appendix.

7 Conclusions

We present ViscoNet, a pioneering approach that seamlessly integrates visual control into a spatial adapter. Our method, characterized by a single branch handling both pose and visual control stands out for its lightweight design and significantly smaller footprint when compared to existing two-adapter solutions. Through a comprehensive blend of qualitative and quantitative assessments, we demonstrate the remarkable efficacy of ViscoNet in seamlessly bridging and harmonizing text and visual prompts. This unique capability not only mitigates mode collapse but also empowers the model to excel across diverse tasks, positioning it as one of the most versatile human image generation models available.

Furthermore, our feature masking technique significantly contributes to our model’s strength by preserving the generative power of the backbone image model. Remarkably, this is achieved despite training exclusively on a human image dataset that is orders of magnitude smaller than the datasets used by reference methods. This underscores the efficiency and generalization prowess of ViscoNet in handling image generation tasks with limited training data.

References

1. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)* (11 2023)
2. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *Proceeding of International Computer Vision Conference (ICCV)* (4 2023)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019)
4. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *SIGGRAPH* (1 2023)
5. Chen, H., Zhang, Y., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374* (2023)
6. Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186* (2023)
7. Cheong, S.Y., Mustafa, A., Gilbert, A.: Kpe: Keypoint pose encoding for transformer-based image generation. *British Machine Vision Conference (BMVC)* (3 2022)
8. Cheong, S.Y., Mustafa, A., Gilbert, A.: Upgpt: Universal diffusion model for person image generation, editing and pose transfer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* pp. 4173–4182 (4 2023)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Conference on Neural Information Processing Systems (NeurIPS)* (2021), <https://arxiv.org/abs/2105.05233>
10. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR* (8 2022)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022)
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Conference on Neural Information Processing Systems (NeurIPS)* (2014)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)* (2020), <https://arxiv.org/abs/2006.11239>
15. Houlisby, N., Giurui, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. *ICML* (2019)
16. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117* (11 2023)

17. HuggingFace: openai/clip-vit-large-patch14. <https://huggingface.co/openai/clip-vit-large-patch14> (2011)
18. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023)
19. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2human: Text-driven controllable human image generation. SIGGRAPH (2022)
20. Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: Humansd: A native skeleton-guided diffusion model for human image generation. International Conference on Computer Vision (ICCV) (4 2023)
21. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Divide & bind your attention for improved generative semantic nursing. BMVC (7 2023)
22. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context. European Conference on Computer Vision (ECCV) (5 2014)
23. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. Arxiv preprint: 2310.08579 (10 2023)
24. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. Conference on Neural Information Processing Systems (NeurIPS) (2017)
25. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. Arxiv preprint 2302.08453 (2 2023)
26. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. Proceedings of Machine Learning Research (2021), <https://arxiv.org/pdf/2112.10741.pdf>
27. Pinkney, J.: Stable diffusion image variations. <https://github.com/justinpinkney/stable-diffusion> (2022)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. International Conference on Machine Learning (ICML) (2 2021)
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. Arxiv Preprint: 2204.06125 (4 2022)
30. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. International Conference on Machine Learning (ICML) (2021)
31. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. IEEE Conference of Computer Vision and Pattern Recognition (CVPR) (4 2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (12 2022)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) (5 2015)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)

35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. Arxiv preprint: 2205.11487 (5 2022)
36. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
37. Stability.ai: Stable diffusion 2. <https://github.com/Stability-AI/stablediffusion> (2023)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Conference on Neural Information Processing Systems (NeurIPS) (2017)
39. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (1 2024)
40. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (6 2023)
41. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003). <https://doi.org/10.1109/ACSSC.2003.1292216>
42. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arxiv:2311.16498 (11 2023)
43. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. Arxiv Pre-print 2308.06721 (8 2023)
44. Zhang, J., Li, K., Lai, Y.K., Yang, J.: Pise: Person image synthesis and editing with decoupled gan. IEEE Conference of Computer Vision and Pattern Recognition (CVPR) (3 2021)
45. Zhang, K., Sun, M., Sun, J., Zhao, B., Zhang, K., Sun, Z., Tan, T.: Humandiffusion: a coarse-to-fine alignment diffusion framework for controllable text-driven person image generation. Arxiv Preprint 2211.06235 (11 2022)
46. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. International Computer Vision Conference (ICCV) (2 2023)
47. Zhang, P., Yang, L., Lai, J., Xie, X.: Exploring dual-task correlation for pose guided person image generation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (3 2022)
48. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. NeurIPS (5 2023)
49. Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., Li, Q.: Cross attention based style distribution for controllable person image synthesis. European Conference on Computer Vision (ECCV) IEEE Conference of Computer Vision and Pattern Rec (8 2022)
50. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. IEEE Conference of Computer Vision and Pattern Recognition (CVPR) (4 2019)
51. Ziwei, Luo, P., Qiu, S., Wang, X., Tang, X.L.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

Appendix:

ViscoNet: Bridging and Harmonizing Visual and Textual Conditioning for ControlNet

No Author Given

No Institute Given

This appendix is split into 3 sections:

- **Section A** provides a further quantitative and qualitative comparison with IP-Adapter 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
	<u>CLIP score</u>								
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
	<u>Mode Collapse Rate (MCR)</u>								
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
	<u>MS-SSIM</u>								
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 **51** 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-Adapter. IP-Adapter 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 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-Adapter 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 21 and Figure 22.



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

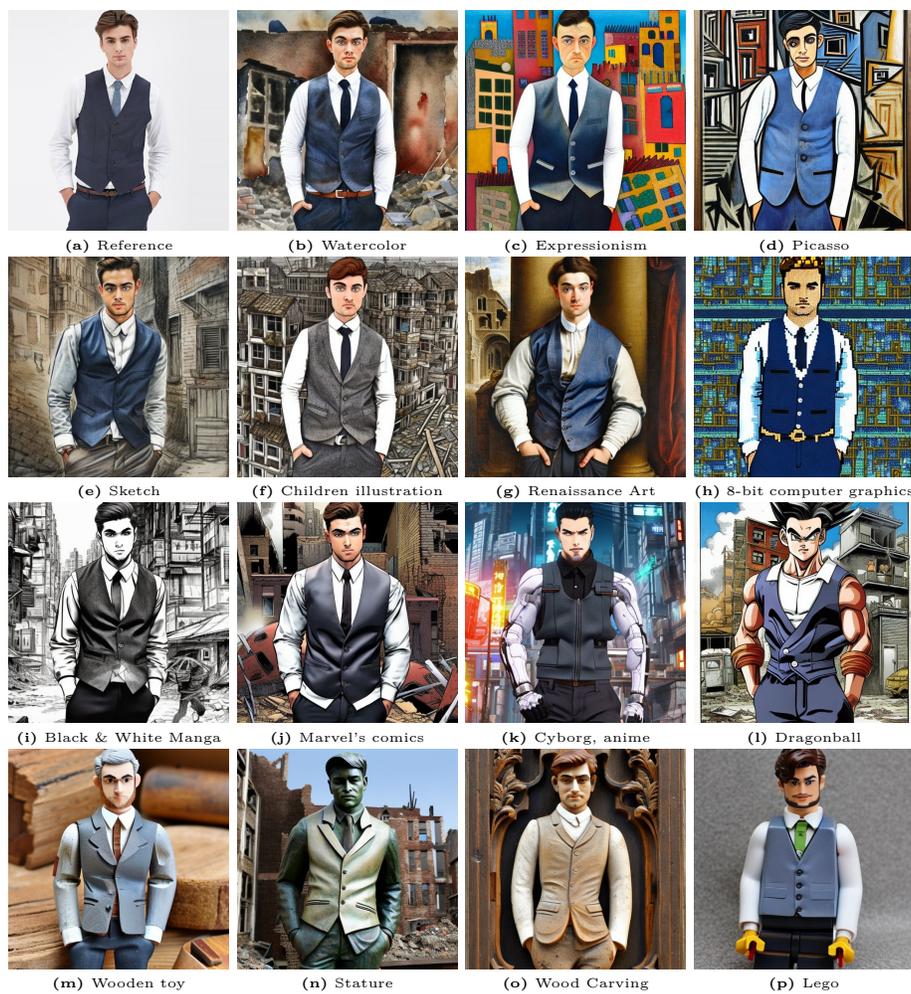


Fig. 22: Stylization. Text prompt: “a man, in a derelict city.”

B.3 Pose Re-target

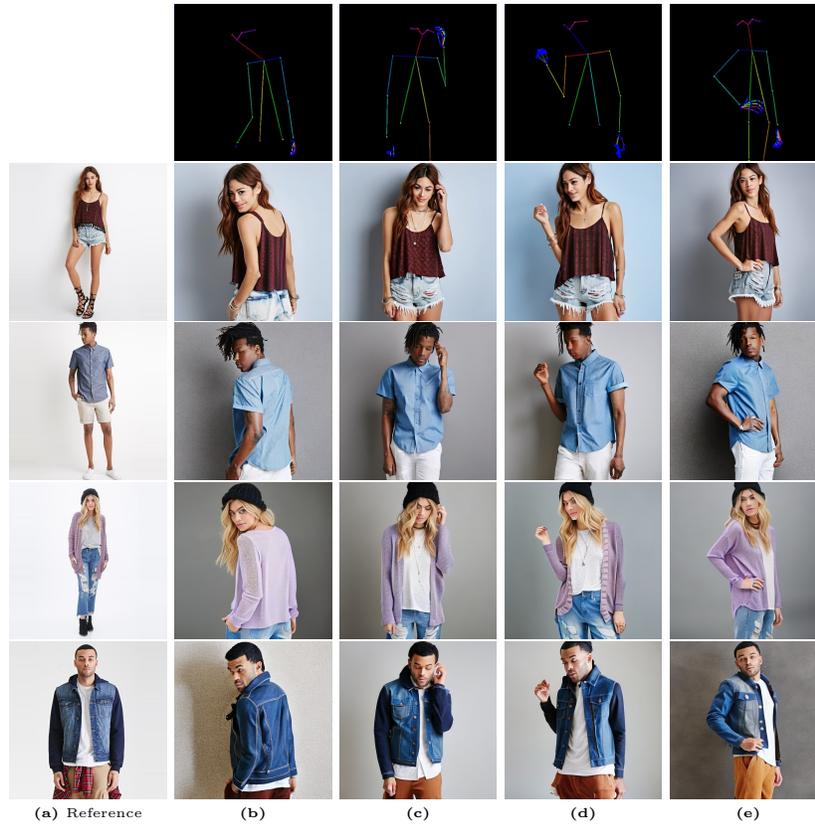


Fig. 23: Pose Transfer from (a) reference person to new poses in (b)-(e)

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.



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

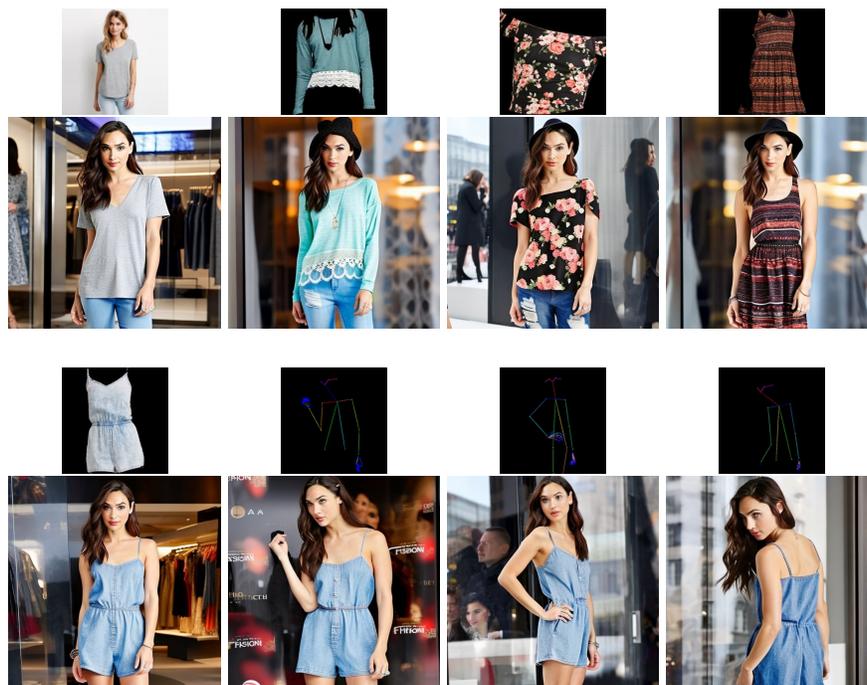


Fig. 25: Combining re-identification, virtual try-on, and pose re-target, we showcase examples of posing fashion with celebrity avatars.

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.

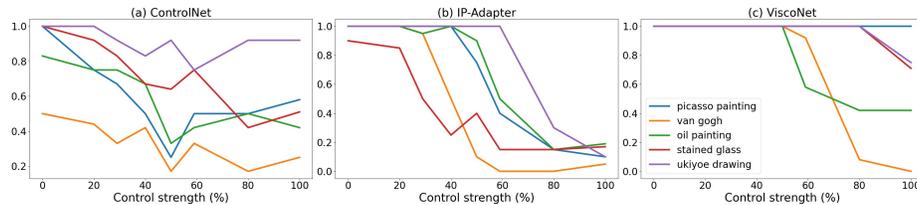


Fig. 26: CLIP accuracy - comparing different image styles.

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.

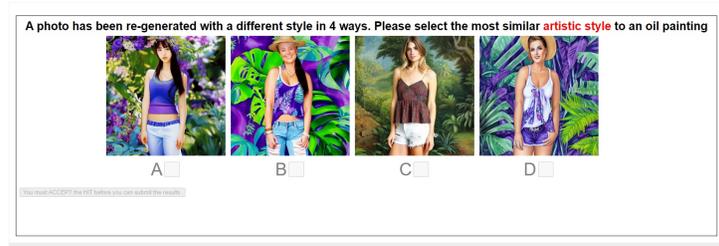


Fig. 27: Screenshot of user study presented to users for evaluating the quality of the stylization against the three baselines.

Image Styles	Human Evaluation				Ours (%)
	HumanSD	ControlNet	T2I-Adapter	ViscoNet (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.