Giving a Hand to Diffusion Models: a Two-Stage Approach to Improving Conditional Human Image Generation

Anton Pelykh, Ozge Mercanoglu Sincan, Richard Bowden

University of Surrey, United Kingdom

{a.pelykh, o.mercanoglusincan, r.bowden}@surrey.ac.uk

| Stable Diffusion [33] | HandRefiner [21] | T2I-Adapter [26] | HumanSD [14] | ControlNet [42] | *Ours* | Pose condition |

Fig. 1: Examples of images, generated by the proposed method (column 6) and the state-of-the-art diffusion models (columns 1 to 5), given the pose condition (final column) and the text description. The text prompts are provided in the supplementary material.

Abstract— **Recent years have seen significant progress in human image generation, particularly with the advancements in diffusion models. However, existing diffusion methods encounter challenges when producing consistent hand anatomy and the generated images often lack precise control over the hand pose. To address this limitation, we introduce a novel approach to pose-conditioned human image generation, dividing the process into two stages: hand generation and subsequent body outpainting around the hands. We propose training the hand generator in a multi-task setting to produce both hand images and their corresponding segmentation masks, and employ the trained model in the first stage of generation. An adapted ControlNet model is then used in the second stage to outpaint the body around the generated hands, producing the final result. A novel blending technique is introduced to preserve the hand details during the second stage that combines the results of both stages in a coherent way. This involves sequential expansion of the outpainted region while fusing the latent representations, to ensure a seamless and cohesive synthesis of the final image. Experimental evaluations demonstrate the superiority of our proposed method over state-of-the-art techniques, in both pose accuracy and image quality, as validated on the HaGRID dataset. Our approach not only enhances the quality of the generated hands but also offers improved control over hand pose, advancing the capabilities of pose-conditioned human image generation. The source code is available here. [1]**

[1] https://github.com/apelykh/hand-to-diffusion

I. INTRODUCTION

Controllable human image generation is an important task in the field of visual content production with applications in advertising, game character creation and E-commerce amongst others. In recent years, diffusion models have overtaken the field with their flexibility and unprecedented quality of results. They dominate over other generative model types such as generative adversarial networks (GAN) and variational auto-encoders (VAE) [7]. Many works have also explored ways to add pose control to diffusion generators [3, 14, 26, 37, 42]. Some approaches [26, 42] add a trainable branch on top of a frozen pre-trained Stable Diffusion (SD) model [33]. Other works [3, 14] propose natively guided diffusion models that receive conditioning in concatenation with the input.

Despite the remarkable results and increased flexibility of pose-guided approaches, diffusion models often struggle to achieve high-quality hand generation, leading to inaccuracies such as extra or missing fingers, distorted hand poses, and the presence of visual artifacts (see Fig. 1). As the human brain is highly attuned to recognizing the details of human anatomy, including the structure and shape of hands, such failure cases are easily spotted and perceived as unnatural and eerie. In addition, modern diffusion generators do not

1

provide precise control over hand pose and struggle to model hand interactions due to possible occlusions and the anatomic complexity of hands [14, 21].

The fact that hands are high-fidelity objects with many degrees of freedom and a high probability of occlusion makes it a challenge to generate them realistically. At the same time, maintaining the model's generalization in terms of appearance and visual style while producing consistent hand anatomy is challenging. Training datasets rarely combine the volume and diversity of samples needed with high curation quality and precise annotation. Publicly available datasets that include annotated hands and hand interactions often lack visual diversity and tend to concentrate only on hands without including the rest of the body [24, 25, 45]. Using such narrow-domain datasets, that do not present high variety in appearance and style for fine-tuning a pre-trained diffusion model, is typically detrimental. It can lead to decreased generality and expressiveness of the generator. This effect is referred to as "catastrophic forgetting" in the literature [14].

Recent works such as HandRefiner [21] and Concept Sliders [8] attempt to fix the quality of hand generation in SD and SDXL [29] respectively. Gandikota *et al.* identify a low-rank direction in the parameter space of a diffusion model that targets hand quality and allow it to be modified via a Concept Slider. On the other hand, in HandRefiner Lu *et al.* propose to repaint hands in the generated images with a depthmap-conditioned ControlNet [42]. While both approaches show improved hand quality, they aim for general visual plausibility and do not allow pose control, which is a paramount factor in numerous application areas of generative models.

This work tackles the problem of high-quality hand generation in diffusion models while aiming to deliver precise control over the pose and preserve generality and high visual controllability. To our best knowledge, it is the first diffusion approach capable of high quality hand generation with pose control. This is achieved by dividing the task into two sub-tasks, namely hand generation and body outpainting around the generated hands. Such an architectural decision is motivated by the idea of decreasing the variability of data that the hand generator needs to learn from and allowing it to favor pose precision and articulation. At the same time, the outpainting stage leverages a conditional diffusion model that is tuned to accommodate complex hand shapes and is capable of synthesizing diverse appearances and styles. The hand generator is trained in a multi-task setting to produce a segmentation mask along with the main denoising objective, enabling precise body outpainting. To bring the two sub-components together in a coherent way and mitigate artifacts on the mask border, we propose a blending approach that utilizes sequential mask expansion.

The contributions of this work are summarized as follows:

- We propose a novel two-stage diffusion-based approach to human image generation that is capable of producing high-quality hands with precise control over their pose.
- We show that conditional diffusion models can be successfully trained in a multi-task setting, predicting both the added noise and the semantic segmentation mask of the generated object.
- We introduce a blending technique that relies on sequential expansion of the outpainting region. It enables harmonious fusion of the both stages of the generation process while ensuring seamless transition between regions and preservation of detail.
- To demonstrate the effectiveness of the proposed solution, we conduct extensive experiments and comparisons with state-of-the-art models measuring pose precision, including a separate evaluation of hand pose, image quality and text-image consistency.

II. RELATED WORK

A. Image Generation with Diffusion Models

Recently, there has been a significant increase of interest in diffusion models from the computer vision community due to their flexibility and high quality of results, that often dominate other generative model types [7]. A notable branch of research in diffusion models is Denoising Diffusion Probabilistic Models (DDPM) [13, 39] that utilize two Markov chains: a forward chain that noises the data, and a reverse chain that recovers data from noise. Ho *et al.* [13] and Dhariwal and Nichol [7] demonstrated the capability of denoising diffusion models to generate high-quality samples unconditionally with Song *et al.* [40] and Nichol and Dhariwal [28] further proposing optimizations to the inference that allow for a significant speed up of the generation process. GLIDE [27] combined a diffusion model with text conditioning by encoding the input prompt into a sequence of embeddings with a transformer, which is then concatenated with the attention context of each layer. Similarly, DALL-E2 [32] and Imagen [34] employed a modified GLIDE architecture to map the the CLIP [30] and T5-XXL encoder [31] embedding space correspondingly into the image space via the reverse diffusion process, generating images that convey the semantic information of the input caption. While early diffusion approaches were performed in pixel space, Rombach *et al.* [33] proposed Latent Diffusion by moving the denoising process to the lower-dimensional latent space of a pre-trained autoencoder, benefiting from perceptual compression of the modeled data and unlocking greater flexibility for solving various image-to-image and text-to-image tasks.

B. Pose-Conditioned Human Image Generation

Although unconditional and text-conditioned approaches can often produce high-quality realistic results, the limited control over generation makes such models unusable for many content production use-cases.

Generative Adversarial Networks (GAN) [9] have been widely used to introduce pose control to image generation. Ma *et al.* [23] utilized explicit appearance and pose conditioning of a two-stage GAN architecture to generate difference maps between the coarsely generated image and the target to ensure faster model convergence. In contrast,

Siarohin *et al.* [38] proposed an end-to-end method that explicitly models pose-related spatial deformations by employing deformable skip-connections. However, their method requires extensive computations of affine transformations to solve pixel-level misalignment caused by pose differences. Zhu *et al.* [44] introduced a progressive scheme that transfers the initial pose to the target through a sequence of intermediate representations using *Pose-Attentional Transfer Blocks*. Subsequently, Zhou *et al.* [43] proposed a cross-attention-based module that distributes features from semantic regions of the source image to satisfy the target pose instead of directly warping the source features. The aforementioned approaches were mostly developed using fashion datasets and/or low-resolution images without account for the hand pose. With this in mind, Saunders *et al.* proposed GAN-based methods [35, 36] for sign language applications that aim to generate fine-grained hand details. Although the approaches explicitly model hands, they can only produce appearances seen during training and do not generalize to out-of-distribution visual conditions.

Diffusion models have seen extensive use for pose-conditioned human image generation. Bhunia *et al.* [4] achieve pose control by concatenating the skeleton condition to the model input. Additionally, the style image features are passed to cross-attention blocks to better exploit the correspondences between the source and target appearances. Building up on Latent Diffusion [33], a number of works [3, 14, 26, 42] extended it to condition the denoising process on various modalities such as human pose keypoints, sketches, edge maps, depth maps, color palette etc. ControlNet [42] introduces a trainable copy of a Stable Diffusion (SD) encoder to extract features from the condition while keeping the base model frozen during training. Similarly, T2I-Adapter [26] uses lightweight composable adapter blocks for condition feature extraction which can be combined for multi-condition setting. In [42] and [26], the features learned by encoders are combined with the features of the frozen backbone model in an additive manner which may provoke trainable-frozen branch conflicts, as discussed in HumanSD [14]. To mitigate this issue, Ju *et al.* do not employ additional encoders and make all the parameters of the underlying Stable Diffusion model trainable, while trying to mitigate the issue of catastrophic forgetting by using the proposed heatmap-guided denoising loss. They achieve pose control by concatenating the skeletal condition with noisy input latents. Similarly, Baldrati *et al.* [3] extend the input of the SD model to include the human pose image and garment sketch, with textual description being encoded with CLIP and passed into the model through a cross-attention mechanism. Notably, most of the recent pose-conditioned approaches to image generation [3, 4, 14, 26, 37] do not include hand keypoints into a skeleton representation and therefore do not offer control over the hand pose. On the other hand, the models that offer such control, e.g. ControlNet, fail to produce realistic and anatomically correct hands. This shortcoming is tackled by our proposed approach.

III. PROPOSED METHOD

A general overview of the proposed framework is shown in Fig. 2. In this work we propose to break the image generation task into two sub-problems: hand generation and body outpainting around the hands. Firstly, the hand image and the corresponding segmentation mask are produced by the diffusion-based generator, guided with the hand pose condition in the form of a keypoint heatmap. It is possible due to the multi-task training setting that we propose for the generator. The obtained results are further resized and aligned with the global body skeleton to serve as input to the outpainting module. The final image is produced during the second stage by outpainting around the generated hands using the ControlNet [42] model. It is guided by the skeleton image and the segmentation mask, obtained at the previous stage. We are able to bring the two stages together in a harmonious way by using the proposed blending strategy with sequential mask expansion. The division into two sub-components intends to decrease the complexity of the task for the hand generator, enabling it to prioritize pose precision and articulation. At the same time, employing a separate model for the outpainting stage allows our system to have greater control over the generated appearance and style via the text prompt.

In Section III-A we outline the Latent Diffusion paradigm that serves as the foundation for our work. Sections III-B and III-C further describe each stage of the generation process in detail, while Section III-D explains the proposed blending technique that unifies both sub-components and helps to produce a coherent output image.

A. Latent Diffusion Models

The idea of Latent Diffusion [33] is to perform the diffusion process in the latent space of a pre-trained autoencoder to decrease the dimensionality of the data and operate on the feature level instead of raw pixels. The input image $I \in \mathbb{R}^{H \times W \times 3}$ is put through an encoder E to obtain its latent representation $x_0 = E(I)$, where $x_0 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4}$. The latents are subsequently corrupted by noise, following the forward diffusion process, a Markov chain of Gaussian transitions:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $t = 1, ..., T$ is the time step that defines the strength of the added noise, β_t is the noise variance and $q(x_t|x_{t-1})$ is the conditional probability of x_t given x_{t-1}. By utilizing the properties of the above process and performing a reparameterization trick, we can obtain x_t from any time step in the closed form:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha_t}}x_0, (1 - \bar{\alpha_t})\mathbf{I}), \quad (2)$$

where $\bar{\alpha_t} = \prod_{i=1}^{t}(1 - \beta_i)$.

Our goal is to restore a clean sample $\hat{x_0}$ from the noise x_T. However, the reverse process $q(x_{t-1}|x_t)$ is intractable in general case, therefore it is approximated with a Gaussian
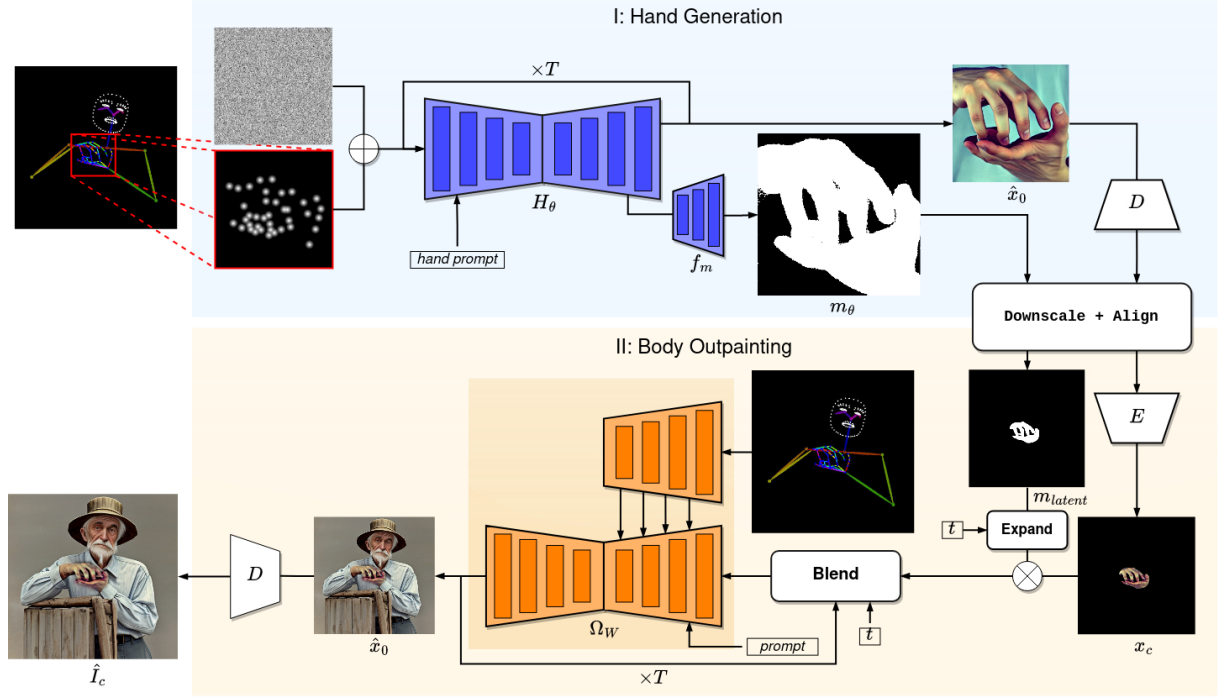
Fig. 2: General overview of the proposed approach. We divide image generation into two sub-tasks: (I) hand generation (top part) and (II) body outpainting around the hands (bottom part).

generative process that utilizes a U-Net model ϵ_θ, trained to predict the added noise and subsequently recover $\hat{x}_0$:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \qquad (3)$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \qquad (4)$$

In DDPM [13] the mean of the reverse diffusion process $\mu_\theta(x_t, t)$ is reparameterized in the following way:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \qquad (5)$$

where $\epsilon_\theta(x_t, t)$ is the predicted noise. DDIM [40] further generalizes DDPM and defines a family of non-Markovian processes with the sampling of a denoised element defined as:

$$p_\theta(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right)$$
$$+ \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(x_t, t), \sigma_t^2\mathbf{I}). \qquad (6)$$

DDIM significantly accelerates the generative process by considering sampling trajectories of length smaller than T without retraining the model.

Latents $\hat{x}_0$, obtained from the denoising process, are then translated back to the pixel space using the decoder D to form the resulting generated image $\hat{I} = D(\hat{x}_0)$.

B. Multi-Task Hand Generation

We use a pre-trained SD model as the foundation for the proposed hand generator H_θ and finetune it in a multi-task setting to predict the noise together with the segmentation mask of the generated hands. Baldrati *et al.* [3] and Ju *et al.* [14] demonstrated that the SD architecture can be successfully conditioned by concatenating additional inputs without employing a separate encoder. Taking inspiration from [3] and [14], our hand generator accepts an additional condition $c_p \in \mathbb{R}^{K \times H_h \times W_h}$ that is concatenated with the noisy input latents, guiding the generation process towards the specified hand shape. The noise term $\epsilon_\theta(x_t, t)$ in (6) is denoted here as $\hat{\epsilon}_\theta$, which is predicted by the proposed conditional model along with the segmentation mask m_θ:

$$\hat{\epsilon}_\theta, m_\theta \leftarrow H_\theta(x_t, t, c_p). \qquad (7)$$

At each step of the diffusion process, a denoised hand latent $\hat{x}_0$ is obtained with DDIM sampling (6) using the predicted noise $\hat{\epsilon}_\theta$.

In this work, a conditional input c_p has $K = 11$ channels, where 10 channels are occupied by the hand keypoint heatmap and 1 channel represents the hand segmentation mask. Each channel of the heatmap contains the keypoints of an individual finger to provide better separability in situations when fingers overlap or are occluded. Furthermore, a segmentation mask of the hands is included in the input to bring additional spatial guidance to the model. During training, the mask is filled with zeros with the probability $p = 0.5$ to increase the robustness of the model and

4

enable mask-free inference. Both the keypoint heatmap and the hand segmentation mask are downsized to the latent dimension with bilinear interpolation to provide explicit pose and layout control to the generator. To accommodate the increased number of input channels, we extend the first convolutional layer of the pre-trained SD architecture with randomly initialized weights and further train the network.

Hand segmentation masks are predicted by a stack of 4 transposed convolutional layers f_m with kernels of the size 2×2 and stride 2. The outputs of each non-final layer are passed through the Sigmoid Linear Unit (SiLU [11]) activation function. The mask prediction head is built on top of the last layer of the SD decoder and it produces outputs in the spatial resolution of the input image I. The predicted mask is further used to define the target region for the body outpainting module and to blend hands and body in a harmonious way.

The network is trained using the combined objective:

$$L = L_{LDM} + \lambda \frac{1}{N} \sum_{i=0}^{N} (M_i - m_\theta)^2, \qquad (8)$$

$$L_{LDM} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t \sim [1,T]} \left[\|\epsilon - \hat{\epsilon}_\theta\|_2^2 \right], \qquad (9)$$

where M_i is the ground truth segmentation mask for the i-th sample, λ is a hyperparameter that defines the weight of the segmentation loss, $\hat{\epsilon}_\theta$, m_θ are the outputs of the hand generator, as shown in (7). Apart from the practical use of the predicted segmentation mask in the next stage, including an extra objective provides an additional regularization to the training process, thus making the generator more robust [10, 16, 20].

C. Body Outpainting

Given the generated hand image and its predicted segmentation mask, the image background is removed. Both the resulting foreground image and the mask are further downscaled and aligned with the full body skeleton to form the canvas for outpainting I_c and its corresponding mask. I_c is then encoded into a latent space using the encoder $x_c = E(I_c)$, and the mask $m_{\theta c}$ is downsized to match the spatial dimensions of the latent representation:

$$m_{\theta c} \in \mathbb{R}^{H \times W} \rightarrow m_{latent} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8}}. \qquad (10)$$

The final generated image is obtained by painting the body around the hand region with a ControlNet model Ω_W. The objective of the model is to predict the unknown latent pixels $(1 - m_{latent}) \odot x_c$ of the input canvas while leaving the masked area $m_{latent} \odot x_c$ unchanged, guided by the body pose in the form of a skeleton image and a mask of the target region. Even though the ControlNet receives the mask as a condition, a diffusion process is performed over the full area of the latent and therefore corrupting the hand region. To preserve the hand details, the latents at each step are obtained by blending the input canvas and the denoised latents at the current step, similarly to [1, 2]:

$$x_t = m_{latent} \odot x_c + (1 - m_{latent}) \odot x_t. \qquad (11)$$

The pre-trained skeleton-conditioned ControlNet model can naturally solve the inpainting task by noising, and subsequently restoring, the masked region of the input. However, in the case of body outpainting, it tends to hallucinate objects and unnatural backgrounds around the hand region as the model learned to associate non-neutral hand shapes to holding objects during the generic training. Furthermore, a pre-trained model often tries to complete the hand outside the bounds of the mask thus making it anatomically incorrect. To mitigate these issues, we fine-tune ControlNet for body outpainting by providing an initial canvas with hands, segmented from the original image, and tasking the model to complete the image by predicting the outside region. We provide the skeleton image as a condition into the model's encoder and blend the noisy and hand latents as described in (11). The masked $L2$ reconstruction loss is used for the training.

D. Sequential Mask Expansion

When outpainting the body around the previously generated hands, it is crucial to ensure hand detail preservation as well as a seamless transition and natural connectivity between the two regions. While the naive blending strategy described in (11) enforces the hand region stays the same throughout the diffusion process, it often leads to anomalies around the region border in the case of non-uniform background. Although tuning the ControlNet for body outpainting helps to alleviate this issue, the model still tends to expand the hand outside the masked region, add extra fingers or introduce erroneous textures for complex hand shapes.

To address the irregularities around the mask border, we propose to gradually dilate the input hand mask for T iterations, where T is the number of diffusion steps, and then use the expanded masks as m_{latent} in (11) starting from the largest and arriving at the original at step T. At the same time, the underlying denoising UNet of the body outpainter receives a precise hand mask at every iteration of the diffusion process. The intuition behind this process is that the possible distortions that the model may manifest around the hand region will be replaced by the latent pixels from the uniform background of the initial canvas. At the same time, the replaced region will be harmonized and blended with the rest of the latent during the next step of diffusion. Using smaller masks for each diffusion step allows washing out the hard border of the expanded region and avoids visible edge artifacts. The last two iterations of diffusion are performed on the full latent without masking to further unify the two regions in terms of transition smoothness, color distribution and shadows.

In Blended Latent Diffusion [1] progressive mask shrinking was employed to enable text-guided image editing in a thin masked region. However, our mask expansion approach is solving a conceptually different task of harmonious blending of two regions of the latent representation with no constraints on the size of the inpainted region. In our case, the diffused region typically spans most of the image and we

5

are forcing it to envelop the generated hands in a coherent and artifact-free way by progressively expanding the mask.

After the diffusion process in completed, the denoised latents are mapped back to the pixel space using the decoder D, i.e $\hat{I}_c = D(\hat{x}_0)$. We then blend the resulting image with the input hand region using the initial mask m_c following the naive strategy from (11). This allows us to reintroduce sharpness to the hands that might have been reduced during the unmasked diffusion steps with no detrimental effects on the blending consistency.

IV. EXPERIMENTS AND RESULTS

A. Datasets

A combination of InterHand2.6M [25], Re:InterHand [24] and HaGRID [15] datasets is used for training the hand gen-erator. The datasets are combined to ensure overall sample quality and diversity. InterHand2.6M is restricted to a studio environment with distinct lighting and a limited number of participants whereas Re:InterHand provides synthetic 3D renders of real images. HaGRID is the most diverse of the three datasets as it was captured "in the wild" but it includes images of varying quality and only bounding boxes as annotation. Both InterHand2.6M and Re:InterHand provide precise hand keypoints and for HaGRID keypoints are extracted using the Mediapipe holistic model [22]. The hand segmentation masks for InterHand2.6M and HaGRID are obtained with SAM ViT-H [17] by using keypoints as queries for the model, while Re:InterHand includes the masks as a part of the dataset. The masks extracted with SAM often include checkerboard artifacts and discontinuities on the edges so they were processed with a 5×5 dilation kernel to mitigate this issue. Fig. 3 shows examples of masks with artifacts and their post-processed versions. We also use the LLaVA-v1.5-7b [19] model to produce image captions for HaGRID. Sequence-level captions for InterHand2.6M and Re:InterHand are created manually to include gender, skin tone and the details of the hand appearance.

To train the hand generator, we crop the square hand regions from the original images and resized them to the

Fig. 3: Segmentation masks extracted with SAM (left), masks after applying a dilation kernel (middle), pixel-wise difference between the two (right)

resolution 512×512 to accommodate the pre-trained SD architecture. For cases where hands are interacting and their bounding boxes intersect, both hands are included in the same crop, otherwise one hand is cropped at random during training. We also apply RGB value shifting and random brightness and contrast changes to augment the training samples. The total training dataset size for the hand generator is $193,300$ samples, where $60,000$ are randomly sampled from InterHand2.6M, $56,500$ from Re:InterHand and $76,800$ from HaGRID training subsets while keeping the original gesture distribution of HaGRID.

To construct the dataset for training the oupainting model, we utilize LAION-Human (from HumanSD [14]). Similarly to how we process HaGRID, the keypoints are extracted with Mediapipe and the hand segmentation masks with SAM ViT-H. Although LAION-Human includes text prompts and pose keypoints, the former are extremely noisy due to the auto-matic way in which the original dataset was constructed, and the latter are too sparse. Therefore, we replace the original text prompts with ones obtained from LLaVA-v1.5-7b, and use Mediapipe estimates as the ground-truth keypoints. The images are filtered to keep only those with a single human present in the frame. We further discard images for which SAM did not produce a hand segmentation mask of at least 2500 pixels. In this way we extract a total of $36,500$ images that are used for model training.

B. Implementation Details

The hand generator is initialized from the official SD v1.5 checkpoint and further tuned for 5 epochs ($30,000$ iterations) on the combination of InterHand2.6M, Re:InterHand and HaGRID, as described in Section IV-A. The segmentation mask loss weight λ from (8) is set to 0.5. The ControlNet model for the body outpainting stage is initialized from the official Openpose-pretrained checkpoint and tuned for 5 epochs (6000 iterations) on our filtered version of LAION-Human. Both models are trained on the Nvidia A1000 GPU with the batch size 32 and learning rate $1e - 5$.

C. Evaluation Metrics

To evaluate the performance of the proposed approach we measure three aspects of the generation: pose accuracy including isolated evaluation of the hand poses, text-image consistency, and image quality. Pose accuracy is measured by Distance-based Average Precision (DAP) [18] and Mean Per Joint Position Error (MPJPE), calculated between the ground truth keypoints and the ones predicted with Mediapipe from the generated images. The core idea behind DAP is to mimic the evaluation metrics for object detection, namely Average Precision (AP) and Average Recall (AR). Originally, AP and AR use the Intersection over Union (IoU) measure for bound-ing boxes, thresholded at different levels, to match the ground truth and predicted objects. In the case of keypoints, IoU is replaced with a distance-based Object Keypoint Similarity (OKS) measure. In addition, MPJPE evaluates the average Euclidean distance between the predicted and ground truth joint positions.
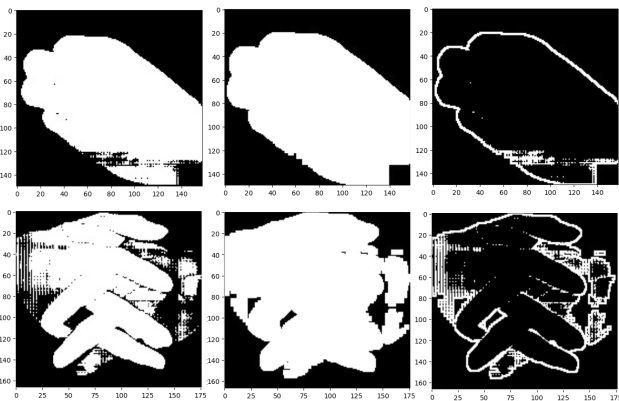
Fig. 4: Visualization of the InceptionV3 convolutional features from the layer with feature dimension 192.

Fréchet Inception Distance (FID [12]) and Kernel Inception Distance (KID [5]) are well established metrics that show the overall quality of the synthesis by comparing the distributions of Inception [41] features extracted from ground-truth and generated images. The FID and KID can be calculated over the features from different layers of the Inception network with the choice of a layer impacting the sensitivity of the metrics to various aspects of image quality and diversity. As this work aims to improve hand generation in diffusion models, we are particularly interested in the quality of hand structure and patters associated with fingers. With this in mind, we explore the Inception features from different layers and identify the feature dimension 192 as the most suitable for our evaluation. We use the Torchmetrics [6] implementation of FID and KID and report the results for the chosen feature dimensionality. The features from a feature dimension 192 are visualized in Fig. 4.

Finally, we use the CLIP [30] similarity score (CLIPSIM) to measure the consistency between the input text prompt and generated images by projecting both into a shared latent space and calculating the distance between the embeddings.

D. Results

The proposed approach is compared to the recent state-of-the-art diffusion-based models, namely SD [33], HandRefiner [21], HumanSD [14], T2I-Adapter [26] and ControlNet [42]. Following the HandRefiner evaluation setting, we randomly sample $12,000$ images from the HaGRID test set, keeping the original gesture distribution, and use them for the comparison. The quantitative results are summarized in Table I.

Firstly, we evaluate the precision of the produced poses for all the approaches by extracting the keypoints from the generated images with Mediapipe and comparing them to the ground truth. We report DAP and MPJPE across all 133 keypoints of the full body (17 for body, 68 for face, 21 for each hand, 6 for feet), as well as separately for 42 hand keypoints. The superiority of the proposed approach in pose controllability is demonstrated by a 30.5% improvement from the baselines for the full body and 92.3% improvement

for the hand DAP. We also outperform the baselines in terms of MPJPE by 50% for the full body and 40% for the hand keypints. It is worth noting that SD and HandRefiner do not allow for pose conditioning and only base the generation on the text prompt. Text prompt conditioning, being an extremely weak guidance for the pose, results in 0.0 DAP. This is because the predicted keypoints are too far from the corresponding ground-truth points for them to be associated with the same body parts by the algorithm.

Initial experiments measuring the image quality showed poor performance despite excellent qualitative results. After investigation, it became apparent that samples from HaGRID often suffer from severe background clutter (see Fig. 5). In the same way the Inception convolutional features, used for FID and KID computation, are sensitive to hand structure, they are also sensitive to clutter in the background. This sensitivity introduces noise to the evaluation metrics. With this in mind, we segment the background out using SAM to ensure a more fair and targeted evaluation of human generation quality. "FID fg" and "KID fg" in Table I report the results on the images with background removed. In this human-centric setting, the proposed approach outperforms the baselines with a 22% improvement. Similarly, our model shows higher results in text-to-image consistency.

E. Ablation Study

It is crucial to employ a reliable blending strategy to combine the results of the hand generator and body outpainter in a harmonious and coherent way. To demonstrate the efficiency of the sequential mask expansion strategy, proposed in Section III-D, we compare it to two alternative approaches: (1) *bounding box blending* and (2) *naive blending*. (1) defines the area outside the square hand region on the canvas as the outpainting region whereas (2) creates the outpainting mask by simply inverting the segmentation mask predicted by the hand generator. In all three cases, the last two steps of the diffusion process are performed with a full mask to smoothen the transitions between the regions. To compare the blending approaches, we randomly sample 500 images from the HaGRID test set, following the original gesture distribution, and measure FID, DAP and MPJPE between the generated images and the originals. It can be seen from Fig. 6 that all three strategies are able to blend the hand and the body coherently. However, (1) does not allow to fully wash out the bounding box region and causes discoloration around

Fig. 5: Examples of images from the HaGRID dataset with severe background clutter.

TABLE I: Quantitative evaluation results on HaGRID dataset

Method	Pose Accuracy				CLIPSIM ↑	Image Quality	
	DAP ↑	DAP hands ↑	MPJPE ↓	MPJPE hands ↓		FID fg ↓	KID fg ↓
Stable Diffusion [33]	0.00	0.00	0.381	0.469	32.94	2.40	1.28 ± 0.30
HandRefiner [21]	0.00	0.00	0.380	0.466	32.95	2.33	1.17 ± 0.28
T2I-Adapter [26]	0.06	0.11	0.179	0.216	33.09	2.39	1.29 ± 0.27
HumanSD [14]	0.31	0.04	0.121	0.236	32.80	2.82	1.58 ± 0.29
ControlNet [42]	0.59	0.39	0.094	0.135	32.86	2.34	1.46 ± 0.34
Ours	**0.77**$_{30.5\%\uparrow}$	**0.75**$_{92.3\%\uparrow}$	**0.047**$_{50\%\uparrow}$	**0.081**$_{40\%\uparrow}$	**34.01**	**1.81**	**0.75 ± 0.045**

TABLE II: Ablation on the blending strategy.

Method	FID ↓	DAP ↑	MPJPE ↓
Bounding Box blending	16.46	0.49	0.087
Naive blending	13.03	0.58	0.062
Sequential Mask Expansion	**12.13**	**0.59**	**0.057**

the hands, corruption of the head and face, if hands are located in close proximity, and "boxy" background artifacts. At the same time, (2) tends to produce anomalies on the border of the hand region that include erroneous extensions of the hands, handheld objects and hallucinated textures. The proposed blending strategy allows us to preserve the area around the hand and eliminate artifacts on the border of the outpainted region. The numerical evaluation results in Table II further demonstrate the sequential mask expansion mechanism outperforming the alternatives in terms of both quality and pose precision metrics. Please refer to the supplementary material for additional qualitative comparison of the three blending methods.
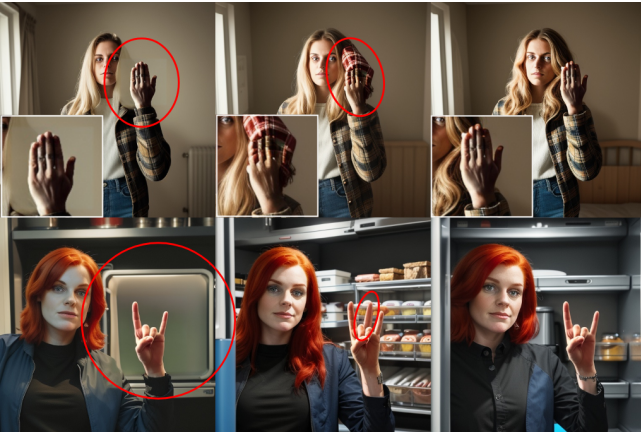

Fig. 6: Qualitative comparison of three blending methods: bounding box (left), naive (middle) and sequential mask expansion (right).

V. LIMITATIONS AND CONCLUSIONS

In this work we presented a novel approach to human image generation that addresses the issue of low-quality hand synthesis and lack of control over the resulting hand pose. The experimental evaluations on HaGRID dataset showed the increased performance of our approach in terms of both pose precision and image quality, comparing to a number of state-of-the-art diffusion-based approaches to image generation.

Although the proposed model produces impressive visual results, there are some limitations to it. We rely on a connectivity between arms and wrists in the input body keypoints. In cases where the hand keypoints are present but the arm is missing from the skeleton, the model may produce discontinuities in the generated image. This is due to the fact that the hand will be generated in the first stage of the process but the arm may not be present to connect to it.

Furthermore, the presented approach concentrates on cases where hands occupy a substantial area in the frame. This is because the spatial dimensions 64×64 of the SD latent space may be insufficient to accommodate the fine details of small hand masks during the outpainting step. Therefore, in the proposed setting, the quality of small hand regions may decrease.

Currently, the results of the hand generator are decoded to the pixel space to be further encoded into latents again for the outpainting stage. As the VAE latent encoding-decoding procedure is lossy, it may result in a decreased quality of the hand region. It is also less efficient from the inference time standpoint. We leave bringing both stages of the process to a shared latent space for future work.

VI. ACKNOWLEDGEMENTS

REFERENCES

[1] O. Avrahami, O. Fried, and D. Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), jul 2023. 5

[2] O. Avrahami, D. Lischinski, and O. Fried. Blended diffusion for text-driven editing of natural images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18197, 2021. 5

[3] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 3, 4

[4] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 3

[5] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 7

[6] N. S. Detlefsen, J. Borovec, J. Schock, A. H. Jha, T. Koker, L. D. Liello, D. Stancl, C. Quan, M. Grechkin, and W. Falcon. Torchmetrics - measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101, 2022. 7

[7] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[8] R. Gandikota, J. Materzyńska, T. Zhou, A. Torralba, and D. Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023. 2

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 5

[11] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 7

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[14] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 1, 2, 3, 4, 6, 7, 8

[15] A. Kapitanov, A. Makhlyarchuk, and K. Kvanchiani. Hagrid - hand gesture recognition image dataset. *arXiv preprint arXiv:2206.08219*, 2022. 6

[16] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2017. 5

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6

[18] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 6

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 6

[20] M. Long, Z. Cao, J. Wang, and P. S. Yu. Learning multiple tasks with multilinear relationship networks. In *Neural Information Processing Systems*, 2015. 5

[21] W. Lu, Y. Xu, J. Zhang, C. Wang, and D. Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. *arXiv preprint arXiv:2311.17957*, 2023. 1, 2, 7, 8

[22] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 6

[23] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 405–415, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

[24] G. Moon, S. Saito, W. Xu, R. Joshi, J. Buffalini, H. Bellan, N. Rosen, J. Richardson, M. Mallorie, P. Bree, T. Simon, B. Peng, S. Garg, K. McPhail, and T. Shiratori. A dataset of relighted 3D interacting hands. In *NeurIPS Track on Datasets and Benchmarks*, 2023. 2, 6

[25] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 2, 6

[26] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024. 1, 3, 7, 8

[27] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. 2

[28] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. 2

[29] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 7

[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 2

[32] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 7, 8

[34] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[35] B. Saunders, N. C. Camgoz, and R. Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020. 3

[36] B. Saunders, N. C. Camgoz, and R. Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151, 2022. 3

[37] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and W. Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023. 1, 3

[38] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe. Deformable gans for pose-based human image generation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3408–3416. Computer

Vision Foundation / IEEE Computer Society, 2018. 3

[39] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. 2

[40] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 7

[42] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 3, 7, 8

[43] X. Zhou, M. Yin, X. Chen, L. Sun, C. Gao, and Q. Li. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision*, pages 161–178. Springer, 2022. 3

[44] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai. Progressive pose attention transfer for person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2347–2356. Computer Vision Foundation / IEEE, 2019. 3

[45] C. Zimmermann, D. Ceylan, J. Yang, B. Russel, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

Giving a Hand to Diffusion Models: a Two-Stage Approach to Improving Conditional Human Image Generation
Supplementary Material

A. Text Prompts

All text prompts presented here have been automatically extracted from the HaGRID dataset test images using the LLaVA-v1.5-7b captioning model. The images in Fig. 1 of the main paper were generated using the following text prompts:

1) "The person in the image is a woman with dark hair, wearing a striped shirt. She is making a peace sign with her hands, and there is a pink wall behind her. The overall visual style of the image is casual and candid, capturing a moment of the woman's life."
2) "The person in the image is a woman with short hair, wearing a gray shirt and black pants. She is holding up four fingers, possibly to make a statement or express her feelings. The overall visual style of the image is a close-up of the woman, focusing on her facial expression and hand gesture."
3) "The image features a man with a beard and a blue shirt. He is sitting down and holding a cell phone up to his ear. The man appears to be making a funny face, possibly for the camera. The overall visual style of the image is casual and candid, capturing a moment of the man's life."

The additional qualitative examples in Fig.8 were generated using the following text prompts:

1) "The person in the image is an older man wearing glasses and a blue shirt. He is making a peace sign with his hand. The overall visual style of the image is a close-up of the man, focusing on his facial features and hand gesture."
2) "The image features a young woman wearing a white shirt with a cartoon character, specifically a tiger, on it. She is making a "peace" sign with her hand. The overall visual style of the image is casual and informal, with the focus on the woman and her gesture."
3) "The image features a woman with short hair, wearing a gray shirt and a black sweater. She is holding a cell phone to her ear, possibly engaged in a conversation. The overall visual style of the image is simple and straightforward, focusing on the woman and her activity."
4) "The person in the image is a woman wearing a green dress, standing in an office setting. She is making a hand gesture, possibly a peace sign or a rock and roll hand gesture, while looking at the camera. The overall visual style of the image is a close-up of the woman, emphasizing her appearance and the hand gesture she is making."
5) "The person in the image is a man wearing a gray shirt and standing in a room. He is making a hand gesture, possibly a rock sign or a peace sign, while posing for the picture. The room appears to be a living space, with a couch visible in the background. The overall visual style of the image is casual and candid, capturing a moment of the man's expression."
6) "The person in the image is a man wearing a blue sweater and glasses. He is making a peace sign with his hand. The overall visual style of the image is a close-up of the man, focusing on his facial expression and hand gesture. The setting appears to be indoors, possibly in a hallway or a room."
7) "The image features a young woman with short black hair, wearing a purple shirt. She is making a thumbs-down gesture, possibly expressing her disapproval or disagreement with something. The woman is standing in a room, and there is a potted plant in the background, adding a touch of greenery to the scene. The overall visual style of the image is casual and candid, capturing the woman's spontaneous reaction to a situation."
8) "The image features a woman with ""long, curly hair, wearing a red shirt. She is making a ""thumbs up"" gesture with her hand, which is a common way to express approval, agreement, or excitement. The overall visual style of the image is casual and candid, as it appears to be a selfie taken by the woman herself."

B. Additional Qualitative Results

Fig. 7: Additional qualitative comparison of the three blending methods: bounding box (left), naive (middle) and sequential mask expansion (right).

Fig. 8: Additional examples of images, generated by the proposed method (column 6) and the state-of-the-art diffusion models (columns 1 to 5), given the pose condition (final column) and the text description.