Handling the Details: A Two-Stage Diffusion Approach to Improving Hands in Human Image Generation

Anton Pelykh, Ozge Mercanoglu Sincan, *MIEEE*, and Richard Bowden, *SMIEEE*

***Abstract*—There has been significant progress in human image generation in recent years, particularly with the introduction of diffusion models. However, it is challenging for the existing methods to produce consistent hand anatomy, and the generated images often lack precise control over hand pose. To address this limitation, we introduce a novel two-stage approach to pose-conditioned human image generation. Firstly, we generate detailed hands and then outpaint the body around those hands. We propose training the hand generator in a multi-task setting to produce both hand image 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. We introduce a novel blending technique that combines the results of both stages in a coherent way and preserves the hand details. It 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 and YouTube-ASL datasets. 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. We make the code available.**

***Index Terms*—Human image generation, hand generation, generative modeling, diffusion models, deep learning, computer vision.**

I. INTRODUCTION

CONTROLLABLE human image generation is an important task in the field of visual content production. It has 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 approaches such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) [6]. Many works have explored ways to add pose control to diffusion generators [2]–[5, 7, 8]. Despite their impressive capabilities and the flexibility introduced by pose conditioning, diffusion models frequently fail in generating high-quality hand images [1]–[5]. Common issues include anatomical inaccuracies such as extra or missing fingers, distorted poses, and visual artifacts (see Fig. 1). These shortcomings are particularly jarring, as the human brain is exceptionally sensitive to hand anatomy, making even minor errors appear unnatural or unsettling. Furthermore, modern diffusion models lack fine-grained control over hand poses and often struggle to accurately model complex interactions or cases with occlusions, reflecting the inherent anatomical complexity of hands [4, 9].

It is also difficult to ensure the model's generalization in terms of visual appearance and style while keeping anatomy consistent. 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 are limited to isolated hand images, omitting context from the rest of the body [10]–[12]. Such data is poorly suited for fine-tuning pre-trained diffusion models, as its limited variability in appearance and style restricts the generator's expressiveness and generality. This performance degradation, known as "catastrophic forgetting", is a well-documented challenge in the field [4].

Recent works attempt to fix the quality of hand generation in diffusion models [9, 13, 14]. In HandRefiner [9], the authors propose to repaint hands in the generated images with a depthmap-conditioned ControlNet [5]. Gandikota *et al.* [14] identify a low-rank direction in the parameter space of SDXL [15] that targets hand quality, and modify it via a Concept Slider. In HanDiffuser [13], Narasimhaswamy *et al.* use a dedicated encoder to predict pose and body shape parameters from the input text prompt, which then serve as conditioning for an image generator. While these approaches show improved hand quality, they aim for general visual plausibility and do not allow pose control, which is a paramount factor in numerous applications of generative models.

This manuscript addresses the problem of high-quality hand generation with diffusion models, focusing on achieving precise pose control while maintaining generality and visual flexibility. Our approach decomposes the task into two complementary sub-tasks: dedicated hand generation and body outpainting around the hands. This division reduces the data variability the hand generator must learn, prioritizing pose accuracy and articulation. At the same time, the outpainting stage employs a conditional diffusion model adapted to handle intricate hand shapes while generating diverse appearances and styles seamlessly. The hand generator operates in a multi-task framework, simultaneously learning to produce segmentation masks alongside the primary denoising objective, enabling precise body outpainting. To coherently merge the outputs

Anton Pelykh, Ozge Mercanoglu Sincan and Richard Bowden are with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, Guildford, U.K. (e-mail: a.pelykh@surrey.ac.uk).

This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), European Union's Horizon2020 programme ('EASIER' grant agreement 101016982) and the Innosuisse IICT Flagship (PFFS-21-47).

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.

of both stages and minimize artifacts at mask boundaries, we introduce a blending technique that utilizes sequential mask expansion. To the best of our knowledge, this is the first diffusion-based image generation approach to successfully produce high-quality hands with precise pose control.

The main contributions of this work are:

- We propose a novel two-stage diffusion approach to human image generation, 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 based on sequential expansion of the outpainting region, enabling smooth integration of the two stages. It ensures seamless transitions between regions while preserving fine details and visual coherence.
- To validate the effectiveness of the proposed approach, we perform extensive experiments and benchmark it against state-of-the-art models, measuring pose precision, including a separate evaluation of hand pose, image quality and text-image consistency.

A preliminary version of this work appeared in [16]. In this extended version, we utilize Low-Rank Adaptation [17] to efficiently fine-tune the model in the second stage of the generation, minimizing the risk of overfitting. Additionally, we have updated the literature to include more recent relevant work and added an ablation study on the effect of the multi-task training of the hand generator. We have also introduced an additional baseline [2], performed an evaluation on the YouTube-ASL dataset [18], and added a user study. We further added qualitative and quantitative results for challenging two-handed poses and expanded the discussion of our model's limitations. Finally, we have improved the overall presentation and writing of the manuscript.

II. RELATED WORK

A. Image Generation with Diffusion Models

There has been a significant interest in diffusion models from the computer vision community due to their flexibility and the high quality of results. For these reasons they dominate over other generative models [6]. A notable branch of research in diffusion models is Denoising Diffusion Probabilistic Models (DDPM) [19, 20] that utilize two Markov chains: a forward chain that noises the data, and a reverse chain that recovers data from noise. Ho *et al.* [20] and Dhariwal and Nichol [6] demonstrated the capability of denoising diffusion models to generate high-quality samples unconditionally. Subsequently, Song *et al.* [21] and Nichol and Dhariwal [22] proposed inference optimizations that significantly speed up the generation process. GLIDE [23] 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 [24] and Imagen [25] employed a modified GLIDE architecture to map the CLIP [26] and T5-XXL encoder [27] 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.* [1] 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 modelled 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 real use cases.

Generative Adversarial Networks (GAN) [28] have been widely used to introduce pose control to image generation through explicit appearance and pose conditioning [29], spatial deformations [30], pose transfer [31], and cross-attention mechanisms [32]. However, they were mostly developed using fashion datasets and/or low-resolution images without accounting for the hand pose. More recently, Saunders *et al.* proposed GAN-based methods [33, 34] 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. Building up on Latent Diffusion [1], a number of works extended it to condition the denoising process on various modalities such as human pose keypoints, sketches, edge maps, depth maps, colour palette, etc. [2]–[5, 7]. ControlNet [5] 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 [3] uses lightweight composable adapter blocks for condition feature extraction, which can be combined for multi-condition setting. However, in both models, 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 [4]. To address this, Ju *et al.* propose making all the parameters of the underlying SD model trainable, while trying to mitigate the issue of catastrophic forgetting by using the heatmap-guided denoising loss. Other approaches, such as Baldrati *et al.* [7] extend the input to include human pose image and garment sketch, while Wang *et al.* [2] propose a ViT-based [35] self-attention module to focus on only the relevant pose embedding tokens.

Numerous works have explored pose-guided image synthesis given a reference appearance image using diffusion models [8, 36, 37]. Bhunia *et al.* [36] 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. Shen *et al.* [8] propose a three-stage approach to bridge the gap between the source and target poses by predicting the global image features first, generating a coarse image, and then further refining it to add detail and improve consistency. Lu *et al.* [37] implement a coarse-to-fine appearance control method based purely on images instead of relying on text prompts. They introduce a perception-refined decoder to decouple fine-grained appearance and pose controls, and use multi-scale attention to enhance the alignment of the source image with the target pose.

Notably, most of the recent pose-conditioned approaches to image generation [2]–[4, 7, 8, 36, 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 [5], fail to produce realistic and anatomically correct hands. This shortcoming is tackled by our proposed approach.

III. PROPOSED METHOD

An overview of our approach is shown in Fig. 2. In this work, we divide the image generation task into two sequential stages: hand generation and body outpainting around the hands. Firstly, a diffusion-based generator produces the hand image and a corresponding segmentation mask, guided by a keypoint heatmap representing hand pose. This is achieved though the multi-task training setting that we employ for the hand generator. The generated hand image is then resized and aligned with the global body skeleton to serve as input for the second stage. The final image is produced in the second stage by outpainting the body around the generated hands using ControlNet [5]. It is guided by the skeleton image and the segmentation mask obtained on the previous stage. To seamlessly integrate the two stages, we propose a blending strategy based on sequential mask expansion, which harmonizes transitions and preserves visual coherence. This modular design simplifies the task for the hand generator, allowing it to prioritize pose precision and articulation, while the outpainting stage provides greater flexibility in appearance and style through text prompts.

Section III-A outlines the Latent Diffusion paradigm underlying our method. Sections III-B and III-C detail each stage of the generation process, while Section III-D describes the proposed blending technique that unifies both stages into a coherent output.

A. Latent Diffusion Models

The idea of Latent Diffusion [1] 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:

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

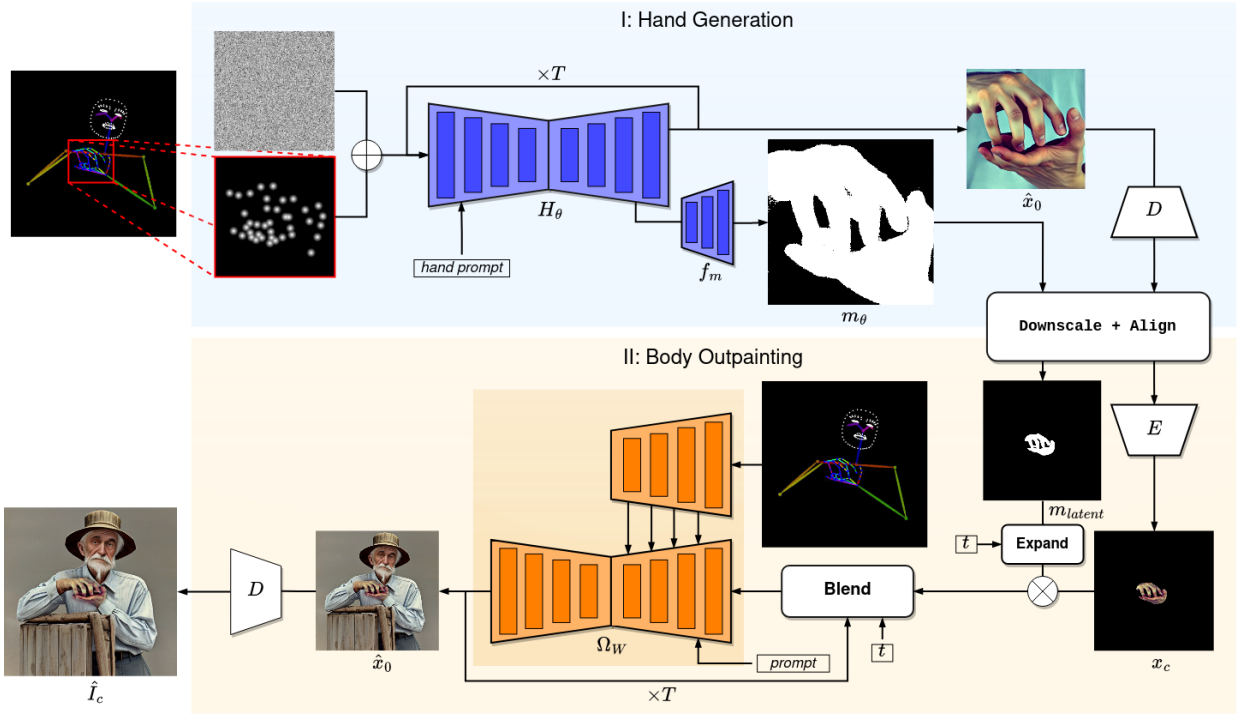

Fig. 2. General overview of the proposed approach. We divide image generation into two sub-tasks: (I) hand generation (top part) and (II) body outpainting around the hands (bottom part).

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.

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 generative process that utilizes a U-Net model ϵ_θ, trained to predict the added noise and subsequently recover $\hat{x_0}$:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{2}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3}$$

In DDPM [20] the mean of the reverse diffusion process $\mu_\theta(x_t, t)$ is reparameterized in the following way:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{4}$$

where $\epsilon_\theta(x_t, t)$ is the predicted noise. DDIM [21] further generalizes DDPM and defines a family of non-Markovian processes that consider sampling trajectories of length smaller than T without retraining the model. This deterministic sampling process allows for fewer denoising steps, significantly speeding up inference.

Latents $\hat{x_0}$, obtained from the denoising process, are then translated back to the pixel space using the decoder D to form the 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. Taking inspiration from [7] and [4], our hand generator accepts an additional input condition $c_p \in \mathbb{R}^{K \times H_h \times W_h}$ that is concatenated with noisy latents, guiding the generation process towards the specified hand shape. The noise term $\hat{\epsilon}_\theta$ is predicted by the model along with the segmentation mask m_θ:

$$\hat{\epsilon}_\theta, m_\theta \leftarrow H_\theta(x_t, t, c_p). \tag{5}$$

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

In our approach, the conditional input c_p has $K = 11$ channels: 10 channels for the hand keypoint heatmap and 1 channel for the segmentation mask of the hand. Each heatmap channel represents the keypoints of an individual finger, ensuring better separability in cases of finger overlap or occlusion. The segmentation mask is included to provide additional spatial guidance to the model. To increase robustness and enable mask-free inference, the mask is set to zeros during training with the probability $p = 0.5$. Both the keypoint heatmap and the segmentation mask are downsized to the latent dimension using 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 intermediate layer are passed through the Sigmoid Linear Unit (SiLU [38]) 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 later used to define the target region for the body outpainting module and to coherently blend hands and body.

The network is trained using the combined objective:

$$L = L_{LDM} + \lambda \frac{1}{N} \sum_{i=0}^{N} (M_i - m_\theta)^2, \tag{6}$$

$$L_{LDM} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t \sim [1,T]} \left[\|\epsilon - \hat{\epsilon}_\theta\|_2^2 \right], \tag{7}$$

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 (5). 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 [39]–[41]. The effect of multi-task training of the hand generator is explored in Section IV-E.

C. Body Outpainting

The background of the generated hand image is removed using the predicted mask. 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 $m_{\theta c}$. The canvas is then encoded into a latent space using the encoder $x_c = E(I_c)$, and the mask is downsized to match the spatial dimensions of the latent representation $m_{\theta c} \in \mathbb{R}^{H \times W} \rightarrow m_{latent} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8}}$.

The final 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. A 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, inexistent hand parts, and unnatural textures around the hand region, as the model learned to associate non-neutral hand shapes with holding objects during the generic training. In the previous work [16], we fine-tuned ControlNet for body outpainting to address these issues. However, direct fine-tuning can negatively affect the generalization ability of the model if performed on a small or noisy dataset. Additionally, the domain gap between generic inpainting and body outpainting is small, which suggests the use of more efficient tuning techniques. Low-Rank Adaptation (LoRA) [17] is a parameter-efficient fine-tuning method that injects trainable low-rank matrices into the original model weights, allowing targeted adaptation without full model updates. In the literature, it was explored for efficient learning of visual concepts and styles in diffusion models [42]–[45], as well as for combining multiple input conditions [46]. In this work, we propose to use a LoRA module on top of a pre-trained ControlNet to specialize it in body outpainting while minimizing the risk of overfitting. Additionally, as the number of trainable parameters becomes $< 1\%$ of the original ControlNet size, the fine-tuning can be performed faster and using only a fraction of GPU memory. The used LoRA configuration and training times of the model are discussed in more detail in Sections IV-E, IV-B. We train the LoRA layers by reconstructing the original image from a canvas containing only segmented hands, thus emulating a two-stage generation process.

D. Sequential Mask Expansion

When outpainting the body around the generated hands, it is essential to preserve hand details while ensuring seamless transitions and natural connectivity between regions. Even though ControlNet receives the mask as an input 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 [47, 48]:

$$x_t = m_{latent} \odot x_c + (1 - m_{latent}) \odot x_t. \tag{8}$$

Although this naive blending strategy keeps the hand region unchanged during diffusion, it often produces border artifacts on non-uniform backgrounds. Fine-tuning ControlNet for body outpainting reduces these issues but does not fully eliminate them, as the model may still expand hands beyond the masked region, add extra fingers, or introduce incorrect textures for complex hand shapes. To address 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. We then use the dilated masks for blending, m_{latent} in (8), starting from the largest and arriving at the original at step T. At the same time, the denoising UNet of the body outpainter receives a precise hand mask at every iteration of the diffusion process. This approach ensures that possible distortions around the hand region are replaced with latent pixels from the uniform background of the initial canvas, while subsequent diffusion steps harmonize and blend the replaced region with the rest of the latent. Using progressively smaller masks for each next diffusion step allows to wash out the border of the expanded region and avoid visible edge artifacts. The last two diffusion iterations are performed on the full latent without masking, further unifying both regions to ensure smooth transitions, consistent color distribution, and realistic shadows.

In Blended Latent Diffusion [48] progressive mask shrinking was employed to enable text-guided image editing in a thin masked region. However, our mask expansion approach solves 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 are forcing it

Fig. 3. Examples generated by our proposed approach (right of each image pair) from the corresponding input pose (left of each image pair). The bottom row shows common failure cases.

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

After the diffusion process is completed, the denoised latents are mapped back to the pixel space using the decoder $\hat{I}_c = D(\hat{x_0})$. We then blend the resulting image with the input hand region using the initial mask m_c following the naive strategy from (8). 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 [10], Re:InterHand [11] and HaGRID [49] datasets are used for training the hand generator. 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. We also curate a partition of the YouTube-ASL [18] dataset, a large-scale open-domain American Sign Language dataset, that we use for model testing. It consists of 2000 video frames that include two-handed signs across a variety of poses and signer appearances, sampled from 50 different videos.

Both InterHand2.6M and Re:InterHand provide precise hand keypoints and for HaGRID and YouTube-ASL keypoints are extracted using the Mediapipe holistic model [50]. The hand segmentation masks for InterHand2.6M and HaGRID are obtained with SAM ViT-H [51] by using keypoints as queries for the model. Masks extracted with SAM often include checkerboard artifacts and discontinuities on the edges, so we process them with a 5×5 dilation kernel to mitigate this issue. We also use the LLaVA-v1.5-7b [52] model to produce image captions for HaGRID and YouTube-ASL. Sequence-level captions for InterHand2.6M and Re:InterHand are created manually to include gender, skin tone and details of the hand appearance.

To train the hand generator, we crop the square hand regions from the original images and resize them to the 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 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 close to $200,000$ samples, where around $60,000$ are randomly sampled from InterHand2.6M, $60,000$ from Re:InterHand and $80,000$ from the training subset of HaGRID, keeping the original gesture distribution.

We utilize LAION-Human (from HumanSD [4]) to construct the dataset for training the outpainting model. Similarly to how we process HaGRID, the keypoints are extracted with Mediapipe and the hand segmentation masks are obtained with SAM ViT-H. We use text prompts from LLaVA-v1.5-7b as ground-truth instead of the original ones provided by LAION-Human due to their noisiness. We extract $36,000$ images, excluding ones with multiple people in the frame and those with hand segmentation masks smaller than 2500 pixels.

B. Implementation Details

The hand generator is initialized from the official SD v1.5 checkpoint and further tuned for 5 epochs (around $30,000$ iterations) on the combination of InterHand2.6M, Re:InterHand and HaGRID, as described in Section IV-A. The segmentation

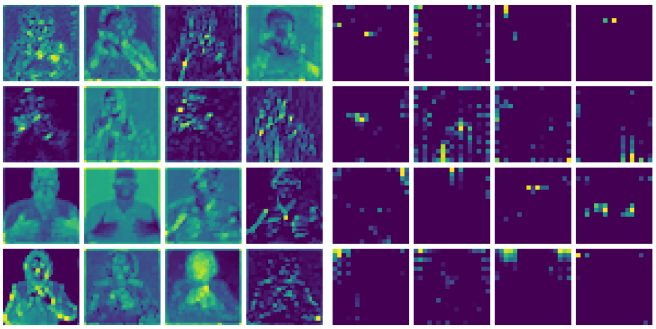

Fig. 4. Visualization of 192-dimensional (left) and 768-dimensional (right) InceptionV3 features.

mask loss weight λ from (6) is set to 0.05. The model is trained on a single Nvidia A100 GPU with batch size 32 and learning rate $1e^{-5}$ over 14 hours. The ControlNet model for the body outpainting stage is initialized from the official Openpose-pretrained checkpoint. LoRA of rank 8 is initialized on top of the "to_k", "to_q", "to_v" and "$to\_out.0$" modules of the ControlNet attention layers, while the rest of the model's weights are kept frozen. It is trained for 5 hours on a single Nvidia RTX3090 GPU with batch size 16 on our filtered version of LAION-Human. We use the masked $L2$ reconstruction loss to train the LoRA layers for body outpainting.

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) [53] and Mean Per Joint Position Error (MPJPE), calculated between the ground truth keypoints and the ones predicted by Mediapipe from the generated images.

Fréchet Inception Distance (FID) [54] and Kernel Inception Distance (KID) [55] are well established metrics that show the overall quality of the synthesis by comparing the distributions of Inception [56] features extracted from ground-truth and generated images. 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. The 2048-dimensional features of the final pooling layer, typically used for the FID/KID evaluation, are designed to represent high-level visual information and may not capture the subtle characteristics of human hands. Alternative layers for FID computation have also been explored in the literature [57, 58]. Motivated by this, we use 192-dimensional features from an intermediate Inception layer for our evaluation, as they are more sensitive to changes in fine image details due to their spatial extent. This adaptation allows for a more relevant evaluation of the model performance in generating high-quality hands, which is central to the objectives of this paper. A visual comparison between the features of dimensions 192 and 768 is shown in Fig. 4.

Finally, we use the CLIP [26] similarity score (CLIPSIM) to measure 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 recent state-of-the-art diffusion-based models, namely SD [1], HandRefiner [9], HumanSD [4], StablePose [2], T2I-Adapter [3] and ControlNet [5]. 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 comparison. The quantitative results are summarized in Table I.

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 is demonstrated by significant improvements over the baselines in both pose precision and image quality metrics. It is worth noting that SD and HandRefiner do not allow for pose conditioning and only base the generation on the text prompt, which provides extremely weak guidance for the pose, resulting in 0.0 DAP. Similarly, our experiments with StablePose revealed significant challenges in generating coherent hand anatomy across samples. As a result, Mediapipe did not detect any hand keypoints in most cases, leading to DAP values close to zero for the hands.

Initial experiments measuring image quality on HaGRID showed poor performance despite excellent qualitative results. After investigation, it became apparent that that dataset suffers from severe background clutter. In the same way the Inception features, used for FID and KID computation, are sensitive to hand structure, they are also sensitive to clutter in the background. To reduce noise in the evaluation metrics, we segment the background out using SAM to ensure a fair and targeted evaluation of human generation quality. "FID fg" and "KID fg" in Table I report the results on images with the background removed. In this human-centric setting, the proposed approach outperforms the baselines with a 37% improvement. Similarly, our model shows higher results in text-to-image consistency.

Although HaGRID offers a diverse set of human appearances, sizes and camera placements, it concentrates on primarily one-handed gestures. To evaluate the model's performance on more challenging two-handed poses that include hand interactions and occlusions, we use a subset of YouTube-ASL, as described in Section IV-A. The evaluation procedure follows the one described for HaGRID. It is worth noting that YouTube-ASL does not suffer from background clutter, so we report the image quality metrics on full images. Table II summarizes the quantitative results of the evaluation, which are consistent with those for HaGRID. Our proposed model outperforms the baselines in both pose accuracy and image quality by a substantial margin.

To measure the model's quality objectively, we conducted a user evaluation. The participants were presented with 20

samples from our test set, where each sample consisted of a reference pose and 5 generated images from HumanSD [4], StablePose [2], T2I-Adapter [3], ControlNet [5] and our proposed model correspondingly. The samples were randomly selected to uniformly cover a set of 10 gestures from HaGRID, skipping trivial and similar gestures. We asked the participants to rank (from best to worst) each image for both pose reproduction accuracy and the overall image quality. The questions were created to cover two crucial aspects of human image generation and align with our quantitative evaluation procedure. We collected the responses from 30 subjects and these are summarized in Fig. 5 where we show the mean rank for each model along with their standard deviations. The data indicates that our proposed model has the highest average rank across both questions. This aligns well with the quantitative results shown in the tables. Furthermore, our model was ranked first in pose precision in 92% of cases and achieved top ranking in visual quality in 64% of cases. ControlNet was ranked first in pose precision in 5% of cases, and T2I-Adapter in 1.5% of cases. Interestingly, despite the fact that StablePose is second to our model quality-wise in the user study, it has the worst quality in the table.

E. Ablation Study

1) Multi-task Training Objective: In addition to serving a practical purpose during the body outpainting stage, predicting segmentation masks along with hand generation can be beneficial for the model's training dynamics. We investigate the effect of adding the mask prediction component to the training loss of the hand generator by training the model on a combination of InterHand2.6M and Re:InterHand, while using HaGRID for validation. The total size of the training dataset for this study is around 110,000 samples. We train the hand generator model for 5 epochs and evaluate the metrics for 360 images of the validation dataset every 2,000 iterations. We measure FID and KID to capture changes in image quality and DAP to track pose precision improvements throughout

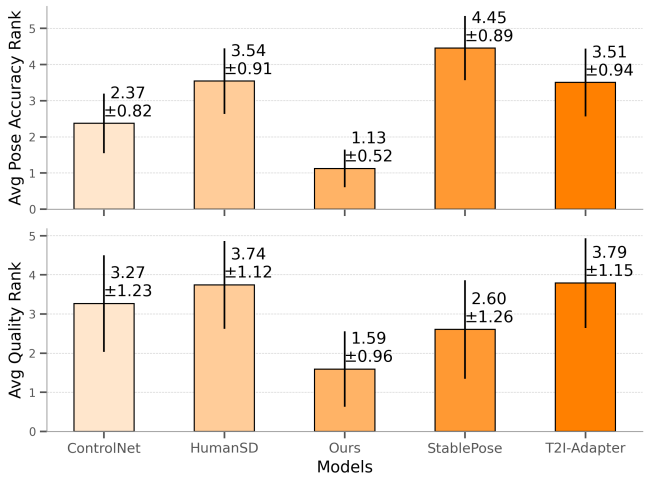

Fig. 5. User evaluation results - mean rank of each model across all the responses with standard deviations.

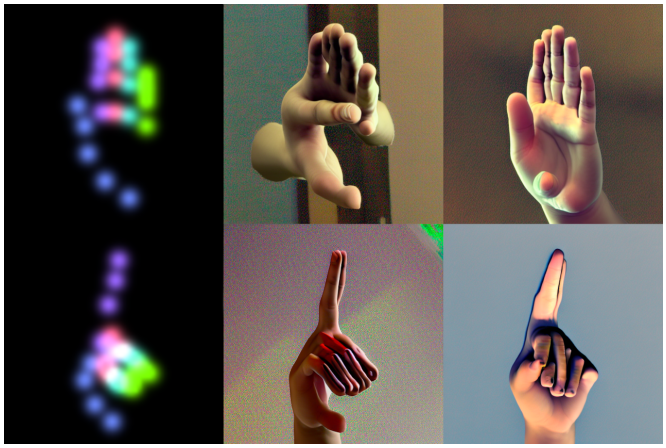

Fig. 6. Comparison of the generated hands on early steps of training without (middle) and with (right) the segmentation mask loss component.

the training. Differently from the main evaluation procedure, FID and KID are calculated using the Inception features of dimension 768. This is because hands occupy most of the frame during the hand generator training, and features closer to the end of the Inception network better capture the global structure of the image unlike previously used features of dimension 192 that concentrate more on local details.

We conduct training without the mask loss, and with two mask loss weights of 0.01 and 0.1 correspondingly. Fig. 7 shows the smoothed plots of the metrics throughout training. As can be seen from FID and KID values, adding a segmentation objective to the training with the weight 0.01 helps the model to converge quicker and achieves better image quality. The higher weight of 0.1 is also beneficial

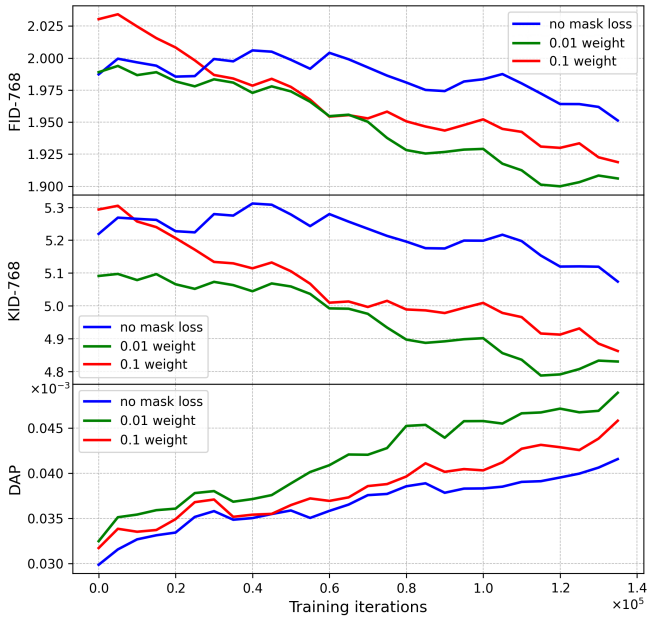

Fig. 7. Image quality (top, middle) and pose accuracy (bottom) evaluation results throughout the training of the hand generator. Blue graphs describe training without the segmentation mask loss, red graphs describe the training with the mask weight 0.1 and green - with 0.01.

TABLE I
QUANTITATIVE EVALUATION RESULTS ON HAGRID DATASET.

Method	Pose Accuracy					Image Quality	
	DAP ↑	DAP hands ↑	MPJPE ↓	MPJPE hands ↓	CLIPSIM ↑	FID fg ↓	KID fg ↓
Stable Diffusion [1]	0.00	0.00	0.381	0.469	32.94	2.40	1.28 ± 0.30
HandRefiner [9]	0.00	0.00	0.380	0.466	32.95	2.33	1.17 ± 0.28
T2I-Adapter [3]	0.06	0.11	0.179	0.216	33.09	2.39	1.29 ± 0.27
StablePose [2]	0.19	0.001	0.178	0.407	33.09	3.67	2.91 ± 0.44
HumanSD [4]	0.31	0.04	0.121	0.236	32.80	2.82	1.58 ± 0.29
ControlNet [5]	0.59	0.39	0.094	0.135	32.86	2.34	1.46 ± 0.34
Ours	**0.76**$_{31\%\uparrow}$	**0.74**$_{89\%\uparrow}$	**0.051**$_{46\%\downarrow}$	**0.088**$_{35\%\downarrow}$	**34.55**	**1.47**$_{37\%\downarrow}$	**0.61 ± 0.048**

TABLE II
QUANTITATIVE EVALUATION RESULTS ON YOUTUBE-ASL DATASET.

Method	Pose Accuracy					Image Quality	
	DAP ↑	DAP hands ↑	MPJPE ↓	MPJPE hands ↓	CLIPSIM ↑	FID ↓	KID ↓
Stable Diffusion [1]	0.00	0.00	0.440	0.622	31.44	32.20	27.57 ± 0.96
HandRefiner [9]	0.00	0.00	0.442	0.628	31.46	33.96	29.85 ± 1.05
T2I-Adapter [3]	0.01	0.02	0.248	0.326	32.13	23.67	21.80 ± 0.92
StablePose [2]	0.13	0.01	0.149	0.219	32.28	44.26	52.91 ± 1.75
HumanSD [4]	0.11	0.01	0.180	0.307	31.79	24.55	23.46 ± 0.94
ControlNet [5]	0.58	0.38	0.110	0.093	32.29	23.62	23.92 ± 1.01
Ours	**0.65**$_{13\%\uparrow}$	**0.7**$_{86\%\uparrow}$	**0.103**$_{6\%\downarrow}$	**0.069**$_{26\%\downarrow}$	**32.95**	**14.94**$_{37\%\downarrow}$	**10.41 ± 0.48**

for performance, compared to the base reconstruction loss. However, the performance increase is smaller as stronger gradients from the auxiliary objective might interfere with the main learning task. This is supported by the DAP values following a similar pattern with the weight 0.01 leading to the best results and having superior training dynamics. We noticed that the mask loss component helps the model to learn a more cohesive structure of the palm during the early steps of the training, as the examples in Fig. 6 show. This may be due to the fact that the input hand keypoint heatmap is sparse in the palm region and the segmentation mask provides an additional spatial context to the model. These results indicate that using a multi-task objective during training improves the performance of the hand generator.

2) Blending with Sequential Mask Expansion: 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. 8 that all three strategies are able to blend the hand and the body coherently. However, (1) does not fully remove the bounding box region and causes discolouration around the hands. Additionally, it corrupts the head and the face, if hands are located in close proximity, and leads to "boxy" artifacts in the background. At the same time, (2) tends to produce anomalies on the border of the hand region that include erroneous extensions of the hands, hand-held 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 III further demonstrate

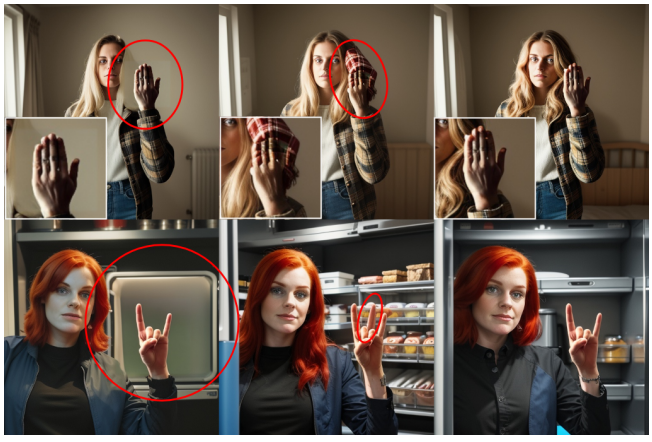

Fig. 8. Qualitative comparison of three blending methods: bounding box (left), naive (middle) and sequential mask expansion (right).

TABLE III
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**

Fig. 9. Examples of generated images from a downscaled input pose by 20%, 40%, 60%, 80% from left to right.

TABLE IV
COMPARISON OF FULL MODEL FINE-TUNING AND LORA.

Method	Pose Accuracy DAP	DAP(h)	MPJPE	MPJPE(h)
Pelykh *et al.* [16]	**0.66**	**0.71**	**0.102**	**0.067**
Ours	0.65	0.7	0.103	0.069

Method	Image Quality FID fg	KID fg	CLIPSIM
Pelykh *et al.* [16]	16.19	12.51 ± 0.46	**33.4**
Ours	**14.94**	$\mathbf{10.41 \pm 0.48}$	32.95

the sequential mask expansion mechanism outperforming the alternatives in terms of both quality and pose precision metrics.

3) Low-Rank Adaptation for Body Outpainting: By updating only the additional low-rank matrices while freezing the original model, LoRA significantly reduces memory consumption and computational overhead compared to full model fine-tuning. The LoRA that we train on top of a ControlNet for the body outpainting stage constitutes around 0.2% of the total number of model parameters. This results in a larger batch size and faster back-propagation, leading to a $5\times$ decrease in training time. Table IV shows a quantitative comparison on YouTube-ASL between the ControlNet fine-tuning proposed in [16] and LoRA. It can be seen that LoRA produces comparable results in pose accuracy, while resulting in increased image quality. This is most likely due to the fact that LoRA improves the model's resilience against overfitting on a small dataset and better preserves its original generalization capabilities.

TABLE V
QUANTITATIVE ANALYSIS OF COMPUTATIONAL EFFICIENCY.

Method	Inference Time (sec/iter.)	Memory (Mb)
SD [1]	2.0	2050
ControlNet [5]	2.65	2800
Ours	9.7	8700

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 pose. The experimental evaluations on HaGRID and YouTube-ASL datasets showed the increased performance of our approach in terms of both pose precision and image quality, compared to a number of state-of-the-art diffusion-based human image generation models.

Although the proposed method produces impressive visual results, it carries certain limitations, as shown in the bottom row of Fig. 3. The presented approach concentrates on cases where the 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. To investigate the quality decrease for small-hand regions, we progressively shrink the input pose for a number of test samples from 20 to 80 percent. The qualitative results are shown in Fig. 9. It can be observed that the hand quality decreases and the number of artifacts around the hand area increases for smaller poses. The quality drop is especially evident for the poses, downscaled by more than 40 percent.

Also, because of the two-stage generation process, there can be a visible difference in skin tone and/or lighting between the hands and the rest of the body. This is due to the outpainting model sometimes being incapable of matching the tone of the hands precisely. Furthermore, since we derive the conditioning for the hand generator only from hand keypoints, it occasionally produces a forearm that does not connect plausibly with the rest of the arm. We leave enhancing the synchronization between the two stages of the generation process to future works. In addition, the two-stage process results in slower inference and higher GPU memory requirements, comparing to SD, as separate models are queried at each stage. We present a quantitative comparison of the inference time and memory consumption between our model, SD and ControlNet in Table V. The measurements were performed on a single Nvidia RTX3090 with no performance optimization techniques applied. Finally, even though the proposed sequential mask expansion reduces the number of blending artifacts, they are still possible around the area of the hand, especially for complex hand shapes or hand-to-face interactions.

REFERENCES

[1] 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, 2022, pp. 10 684–10 695. 1, 2, 3, 7, 9, 10

[2] J. Wang, M. Ghahremani, Y. Li, B. Ommer, and C. Wachinger, "Stable-pose: Leveraging transformers for pose-guided text-to-image generation," *arXiv preprint arXiv:2406.02485*, 2024. 1, 2, 3, 7, 8, 9

[3] 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*, vol. 38, no. 5, 2024, pp. 4296–4304. 1, 2, 3, 7, 8, 9

[4] 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*, 2023, pp. 15 988–15 998. 1, 2, 3, 4, 6, 7, 8, 9

[5] 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*, 2023, pp. 3836–3847. 1, 2, 3, 7, 8, 9, 10

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

[7] 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

[8] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and Y. Wei, "Advancing pose-guided image synthesis with progressive conditional diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024. 1, 3

[9] 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, 7, 9

[10] 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*. Springer, 2020, pp. 548–564. 1, 6

[11] 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. 1, 6

[12] 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. [Online]. Available: "https://lmb.informatik.uni-freiburg.de/projects/freihand/" 1

[13] S. Narasimhaswamy, U. Bhattacharya, X. Chen, I. Dasgupta, S. Mitra, and M. Hoai, "Handiffuser: Text-to-image generation with realistic hand appearances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2468–2479. 1

[14] 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. 1

[15] 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. 1

[16] A. Pelykh, O. M. Sincan, and R. Bowden, "Giving a hand to diffusion models: A two-stage approach to improving conditional human image generation," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–10. 2, 5, 10

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9 2, 5

[18] D. Uthus, G. Tanzer, and M. Georg, "Youtube-ASL: A large-scale, open-domain american sign language-english parallel corpus," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=QEDjXv9OyY 2, 6

[19] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 2256–2265. [Online]. Available: http://proceedings.mlr.press/v37/sohl-dickstein15.html 2

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

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

[22] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8162–8171. [Online]. Available: https://proceedings.mlr.press/v139/nichol21a.html 2

[23] 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. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335086 2

[24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. 3

[25] 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*, vol. 35, pp. 36 479–36 494, 2022. 3

[26] 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*. PMLR, 2021, pp. 8748–8763. 3, 7

[27] 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*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html 3

[28] 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*, 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf 3

[29] 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*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 405–415. 3

[30] 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*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3408–3416. 3

[31] 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*. Computer Vision Foundation / IEEE, 2019, pp. 2347–2356. 3

[32] 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*. Springer, 2022, pp. 161–178. 3

[33] 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*, 2022, pp. 5141–5151. 3

[34] ——, "Everybody sign now: Translating spoken language to photo realistic sign language video," *ArXiv*, vol. abs/2011.09846, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227054383 3

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy 3

[36] 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*, 2023, pp. 5968–5976. 3

[37] Y. Lu, M. Zhang, A. J. Ma, X. Xie, and J. Lai, "Coarse-to-fine latent diffusion for pose-guided person image synthesis," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6420–6429. 3

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

[39] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:206771194 5

[40] M. Long, Z. Cao, J. Wang, and P. S. Yu, "Learning multiple tasks with multilinear relationship networks," in *Neural Information Processing Systems*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:4962395 5

[41] 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*, pp. 7482–7491, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:4800342 5

[42] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu *et al.*, "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15 890–15 902, 2023. 5

[43] L. Li, H. Zeng, C. Yang, H. Jia, and D. Xu, "Block-wise lora: Revisiting fine-grained lora for effective personalization and stylization in text-to-image generation," *CoRR*, vol. abs/2403.07500, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.07500 5

[44] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong, "Mace: Mass concept erasure in diffusion models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268351735 5

[45] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254408780 5

[46] Y. Xu, Z. He, S. Shan, and X. Chen, "CtrloRA: An extensible and efficient framework for controllable image generation," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=3Gga05Jdmj 5

[47] 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)*, pp. 18 187–18 197, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244714366 5

[48] O. Avrahami, O. Fried, and D. Lischinski, "Blended latent diffusion," *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023. [Online]. Available: https://doi.org/10.1145/3592450 5

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

[50] 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*, vol. abs/1906.08172, 2019. [Online]. Available: http://arxiv.org/abs/1906.08172 6

[51] 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*, 2023, pp. 4015–4026. 6

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

[53] 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 *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693. Springer, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48 7

[54] 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*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6629–6640. 7

[55] 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. [Online]. Available: https://openreview.net/forum?id=r1lUOzWCW 7

[56] 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)*, 2015, pp. 1–9. 7

[57] E. J. Nunn, P. Khadivi, and S. Samavi, "Compound frechet inception distance for quality assessment of GAN created images," *CoRR*, vol. abs/2106.08575, 2021. [Online]. Available: https://arxiv.org/abs/2106.08575 7

[58] M. Woodland, A. Castelo, M. Al Taie, J. Albuquerque Marques Silva, M. Eltaher, F. Mohn, A. Shieh, S. Kundu, J. P. Yung, A. B. Patel, and K. K. Brock, "Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part XII*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 87–97. [Online]. Available: https://doi.org/10.1007/978-3-031-72390-2_9 7

Anton Pelykh received a BSc in Applied Mathematics and an MSc in Cybersecurity in 2018 from National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". In 2019, he was awarded the Chevening scholarship to pursue an MSc in AI at Queen Mary University of London, which he obtained with distinction in 2020. Anton joined the Cognitive Vision Group within the Centre for Vision, Speech and Signal Processing in the University of Surrey in 2022 to pursue his PhD in computer vision. His research focuses on human generation in different modalities, including images, videos and 3D, for producing content in sign languages.

Özge Mercanoğlu Sincan is a Research Fellow in computer vision and deep learning at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey. Before joining CVSSP, she received her PhD degree from Ankara University, Türkiye, in 2021. Her research interests primarily lie in the area of computer vision and deep learning, with a particular focus on sign language research.

Richard Bowden is Professor of computer vision and machine learning at the University of Surrey where he leads the Cognitive Vision Group within the Centre for Vision, Speech and Signal Processing. His research centres on the use of computer vision to locate, track, and understand humans. He was Associate Editor for Image and Vision Computing 2009-2022 and IEEE Pattern Analysis and Machine Intelligence 2013-2019. He is cofounder of the company Signapse AI. He was a member of the British Machine Vision Association (BMVA) executive committee and a company director for seven years. He is a fellow of the Higher Education Academy, a senior member of the IEEE, a Fellow of the International Association of Pattern Recognition and Distinguished Fellow of the BMVA.