# Supplementary Materials: HandRefiner

Anonymous Authors

## 6.1 Details for Comparison with Stable Diffusion Baseline

For HAGRID evaluation, in the experiment with ‘Gesture Prompt’, we use the prompt “a/an {ethnicity} {gender} facing the camera, making a/an {gesture}, indoor", where “{gender}" can be man or woman, “{ethnicity}" can be asian, black or white, and they follow the frequency distribution of gender and ethnicity in HAGRID. For “{gesture}" we use the following category-to-prompt mapping: “call": “gesture of calling", “dislike": “thumbs down gesture", “fist": “fist", “four": “hand gesture of four", “like": “thumbs up gesture", “mute": “hush gesture", “ok": “ok gesture", “one": “hand gesture of one", “palm", “hello gesture", “peace": “victory hand sign", “peace inv.": “victory hand sign", “rock": “rock and roll gesture", “stop": “stop hand sign", “stop inv.": “stop hand sign", “three": “hand gesture of three", “three 2": “hand gesture of three", “two up": “hand gesture of two", “two up inv.": “hand gesture of two". In the experiment without ‘Gesture Prompt’, we use the same prompt except the “{gesture}" is replaced with generic “hand gesture".

For FreiHAND evaluation, since the hands in the FreiHAND dataset have light skin tones, and most hand gestures are free poses that can’t be described by gesture categories, we use the following prompt “an/a asian/white man/woman making a hand gesture, indoor". We crop around the hand regions in each generated image so that each crop has a centered hand.

For the user survey, we use the prompt “a/an asian/black/white man/woman making a hand gesture" to ensure the diversity of characters. So different hand shapes and skin tones are covered in the generated images.

## 6.2 Additional Details for Method

As a boundary case omitted from the main method section, we also apply the same masking strategy to the initial random initialized noise vector $x_{\text{noise}}$ similar to Eq. 2:

$$x_{\tau_T} = m \odot x_{\text{noise}} + (1 - m) \odot x_{\tau_T}^{\text{known}}, \tag{5}$$

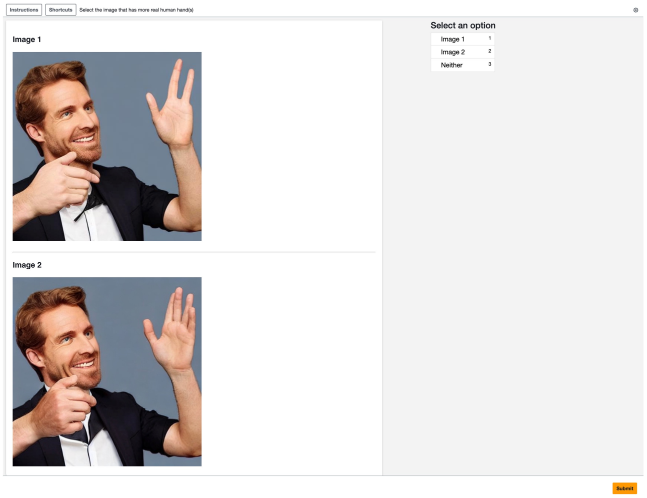
So $x_{\tau_T}$ is the starting noise vector for the iterative sampling procedure.

## 6.3 Procedure for MPJPE Calculation

Given the 3D hand mesh vertices $V \in \mathbb{R}^{778\times3}$ fitted on a malformed hand, we use a sparse linear regressor $\mathbf{J}$ [27] to obtain the 3D coordinates of 21 hand keypoints. We project these keypoints onto image space through a projective function $\Pi$ using a pinhole camera model, resulting in 2D ground truth keypoints $K = \{(x_1, y_1), \ldots(x_{21}, y_{21})\}$. Then, after the conditional generation of the rectified hand, we apply the same mesh reconstruction model again on the rectified hands, and obtain another set of keypoints $K' = \{(x_1', y_1'), \ldots(x_{21}', y_{21}')\}$ using the same procedure. Then MPJPE can be computed by comparing $K$ and $K'$. The process can be formulated as below:

$$K = \Pi(\mathbf{J}V) \tag{6}$$

$$K' = \Pi(\mathbf{J}V') \tag{7}$$



Select the image that has more real human hand(s)

**Figure 1: Screenshot of a survey question in the user survey 4.2**

$$MPJPE(K, K') = \frac{1}{21}\sum_{i=1}^{21} \left\| K_i - K_i' \right\| \tag{8}$$

MPJPE for an image then can be computed by averaging the errors of all hands present in the image.

## 6.4 Example of a survey question

Figure 1 shows a screenshot of a survey question. Note the positions of images are randomized between the original set and the rectified set.

## 6.5 Additional Experiment Results

| Fine-tuned | Control Strength | Negative Prompt | Inpainting Loss | FID ↓ | MPJPE ↓ |
|---|---|---|---|---|---|
| ✗ | 1.0 | ✗ |  | 23.470 | 14.328 |
| ✓ | 0.55 | ✓ | ✓ | 13.977 | 7.878 |

**Table 4: Results of an additional baseline**

In the same setting of Ablation Studies 4.3, we also measure the performance of the HandRefiner using un-finetuned depth ControlNet with control strength set to 1.0 (instead of 0.55 as in the $1^{st}$ row of Table 3). This can serve as an additional baseline, given that no specific techniques have been applied to the model. Table 4 shows its comparison with the finetuned model employing the fixed strength strategy. It demonstrates the finetuned model still greatly outperforms the baseline in both visual quality and pose accuracy, *i.e.*, 9.493 FID and 6.450 MPJPE improvement. This further highlights the effectiveness and necessity of the range of techniques adopted to improve hand generation.

## 6.6 Further Examples of Phase Transition

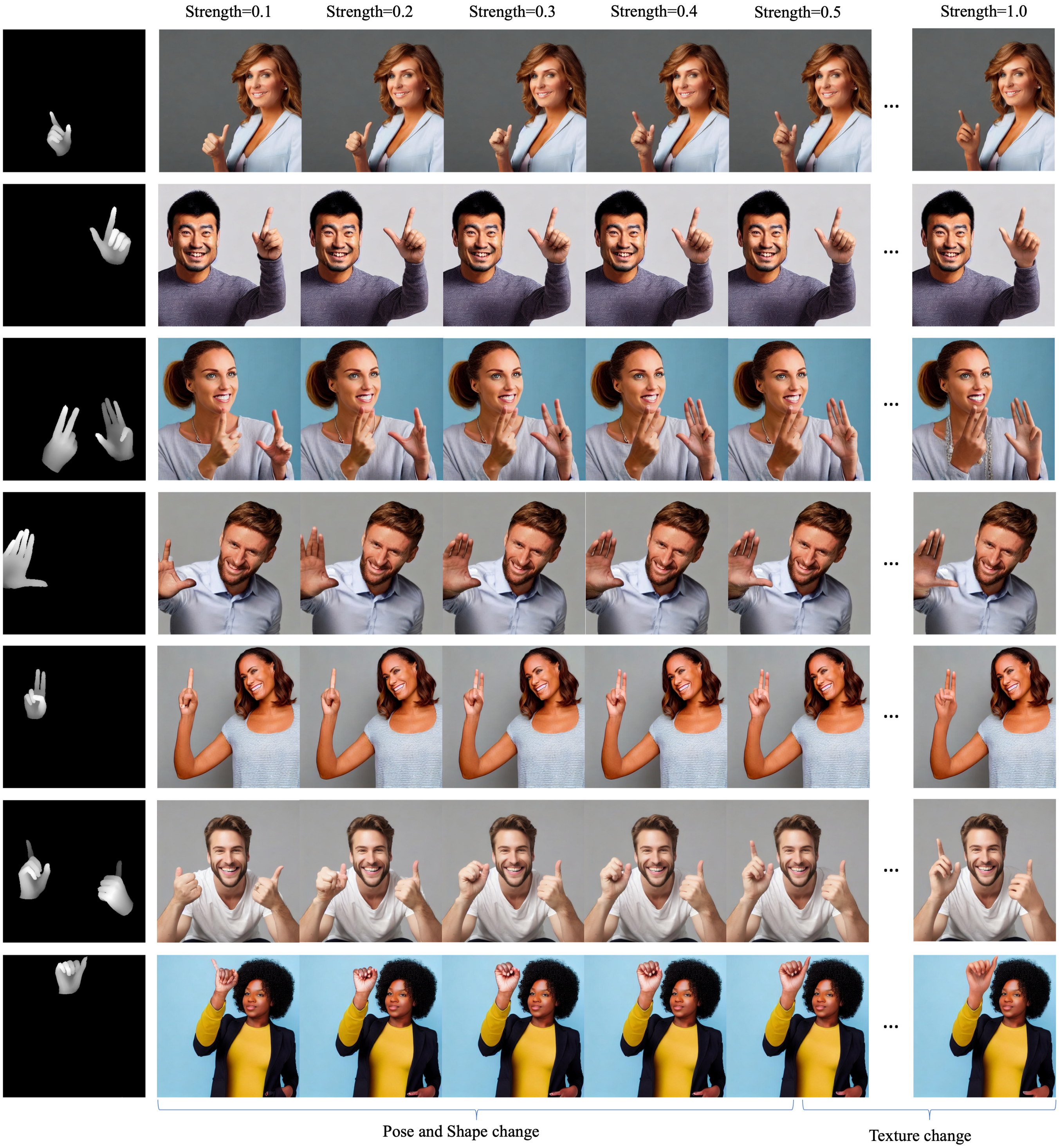
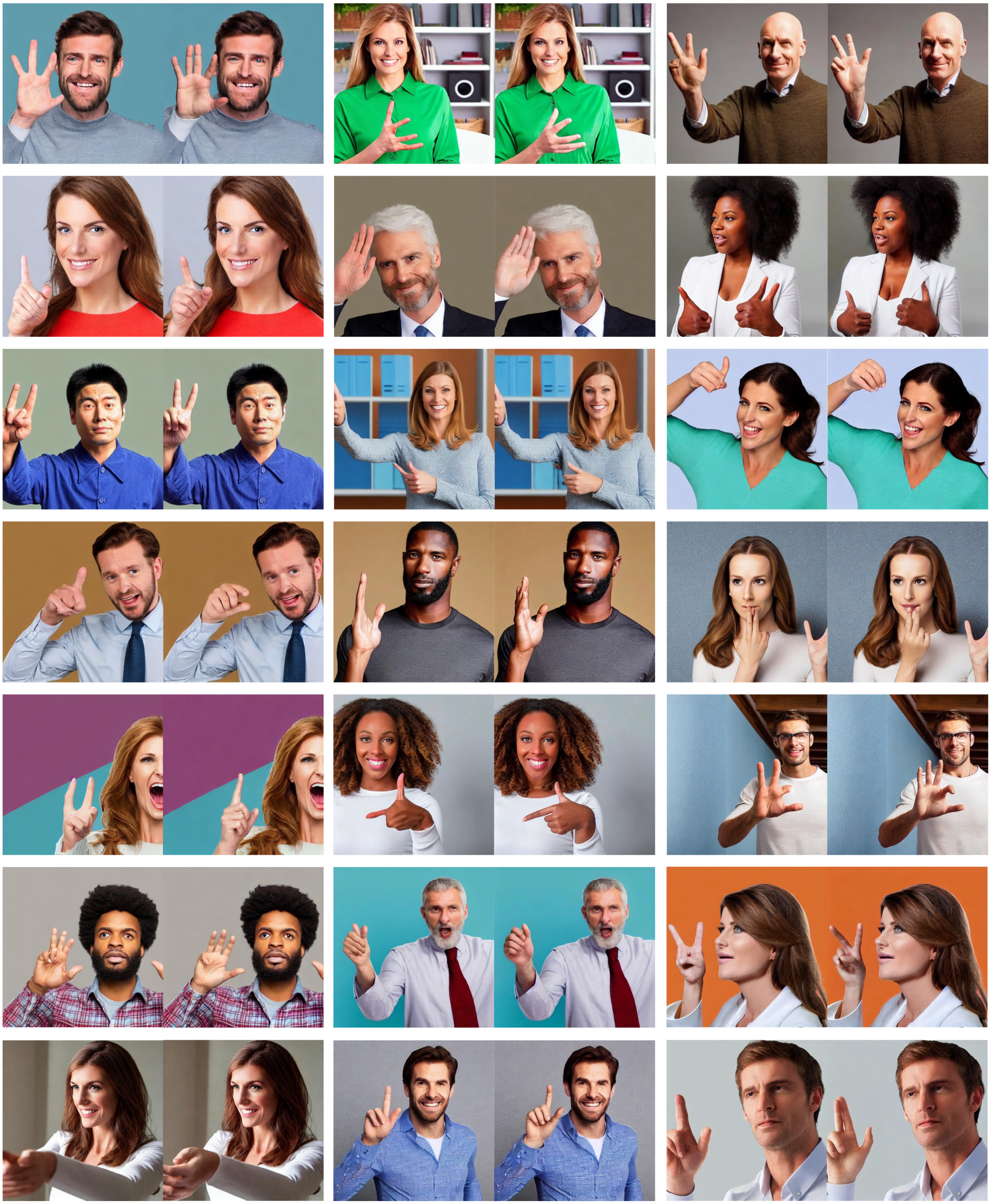


**Figure 2: Illustration of Phase Transition. The hand shape and pose change to match the depth map when applied control strength is small, while the hand wrinkles and textures disappear at high control strength.**

## 6.7 Further Rectification Examples



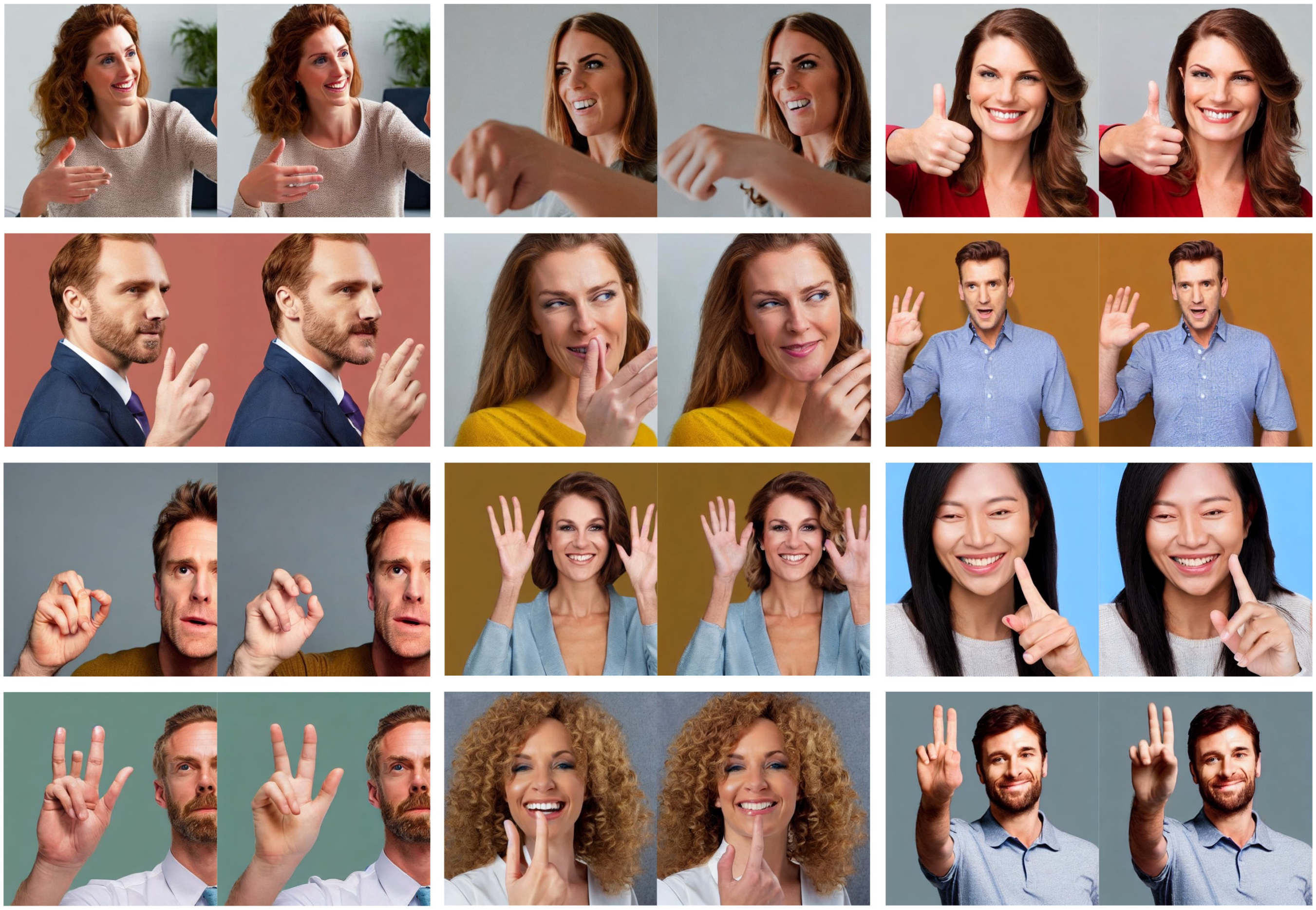


Figure 3: Stable Diffusion generates malformed hands (left in each pair), e.g., incorrect number of fingers or irregular shapes, which can be effectively rectified by our HandRefiner (right in each pair).

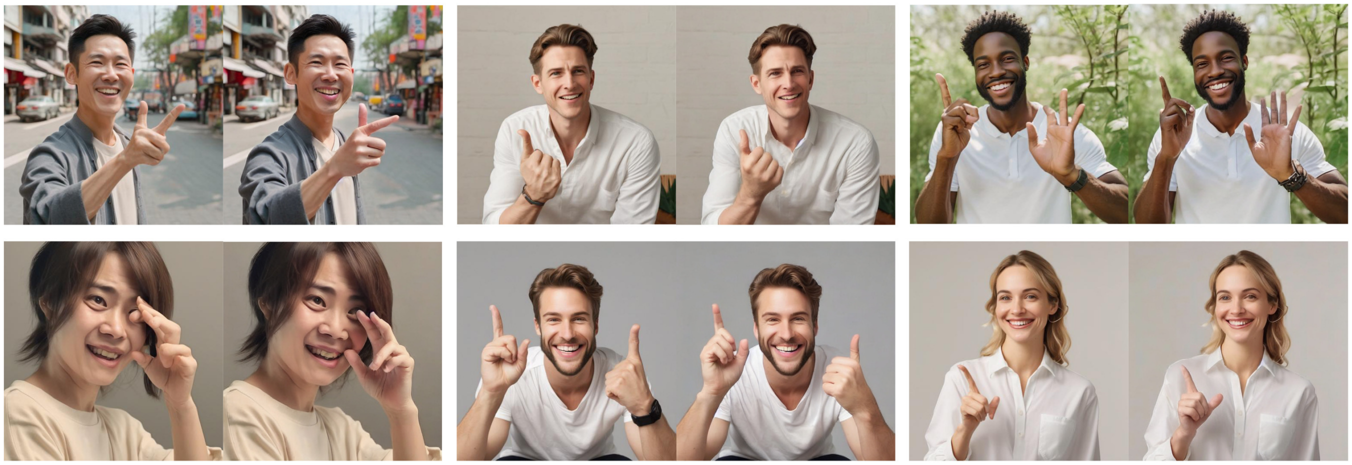


Figure 4: SDXL generates malformed hands (left in each pair), e.g., incorrect number of fingers or irregular shapes, which can be effectively rectified by our HandRefiner (right in each pair).

## 6.8 Depth Map Rendering

To magnify the depth signal, We apply normalization to the depth map of each hand in both training sets and the HandRefiner pipeline. On a scale of 0 (black) to 1 (white), the closest surface is given the value of 1.0, and the furthest surface is given the value of 0.2.