

Hand1000: Generating Realistic Hands from Text with Only 1,000 Images

Haozhuo Zhang¹, Bin Zhu^{2*}, Yu Cao², Yanbin Hao³

¹Peking University

²Singapore Management University

³University of Science and Technology of China

In this supplementary material, we first present a comparative analysis of image captions generated using BLIP2 (Li et al. 2023), PaliGemma (Beyer et al. 2024), and VitGpt2 (Mishra et al. 2024), followed by the results of post-processing these captions with LLAMA (Touvron et al. 2023). Subsequently, we perform a detailed comparison of various hand gesture images produced by Stable Diffusion (baseline) (Rombach et al. 2022) and our Hand1000 model. The results clearly demonstrate that our model outperforms the baseline in handling complex hand gestures: the generated hands not only appear more realistic, but other elements in the images—such as characters, clothing, and colors—also accurately reflect the descriptions in the text.

Dataset Construction

We employed BLIP2, PaliGemma, and VitGpt2 as image captioning models to generate textual descriptions for images in the Hagrid dataset. As shown in Figure 1, BLIP2’s descriptions effectively capture key details of the images, such as characters, clothing, and colors—for instance, “in a green shirt and glasses”—which is beneficial for model training. PaliGemma, however, tends to overemphasize background information while providing less accurate descriptions of the characters themselves. Although VitGpt2 avoids unnecessary details, its descriptions of characters are less precise than those generated by BLIP2.

Additionally, none of the models generated accurate hand gesture information, or they included erroneous gesture details, which could severely hinder training. As is shown in Figure 2, to address this, we employed the LLAMA model for post-processing the generated captions, ensuring the preservation of original content while incorporating correct gesture information. Our experiments demonstrate that training on post-processed data leads to significant improvements in model performance.

Qualitative Example Comparison

To thoroughly evaluate the capabilities of Hand1000, we selected six gestures for testing: ‘call,’ ‘four,’ ‘like,’ ‘mute,’ ‘ok,’ and ‘palm.’ These gestures were chosen due to their diversity and complexity. Specifically, the gestures range from those resembling a fist, such as ‘call,’ ‘like,’ and ‘mute,’ to

| Image | BLIP2 | PaliGemma | VitGpt2 |
|---|---|---|--|
|  | a girl in a green shirt and glasses giving the peace sign | In this image we can see a woman standing and holding an object. In the background we can see a clock, a calendar, a map, a plant in a pot, a table, a kettle and a few other objects | a girl is smiling and holding a piece of food |
|  | a woman in a white sweater making the sign for three | In this image we can see a person wearing black color jacket and white color t -shirt is standing and in the background of the image there is a curtain, air conditioner and there is a light | a woman standing in front of a white wall |
|  | a woman standing in a room with a wall paper | In this image we can see a woman standing and in the background we can see the wall | a woman standing in front of a wall with a scarf around her neck |
|  | a man in a t- shirt and shorts is making a hand gesture | In this image I can see a person standing. In the background I can see a sofa, a table and a curtain | a man standing in a living room holding a cell phone |
|  | a man making a hand gesture in a room | In this image we can see a person standing and giving a pose. In the background there is a wall and a bed | a man standing in front of a mirror in a room |
|  | a woman in a pink shirt making a hand gesture | In this image we can see a woman. In the background there is a wall, door , television, wires and a light | a young girl standing in front of a tv |

Figure 1: Comparison of dataset constructed by different image captioning models (BLIP2, PaliGemma, and VitGpt2).

those with extended fingers, such as ‘four’ and ‘palm,’ as well as the intricate gesture ‘like.’ They also encompass gestures depicting the back of the hand, such as ‘call’ and ‘like,’ and those depicting the palm, such as ‘four’ and ‘palm.’ Ad-

*Corresponding author.

| Image | BLIP2 | BLIP2(post-processed) |
|--|---|--|
|  | a girl in a green shirt and glasses giving the peace sign | a girl in a green shirt and glasses making phone call hand gesture |
|  | a woman in a white sweater making the sign for three | a woman in a white sweater making four fingers up hand gesture |
|  | a woman standing in a room with a wall paper | a woman standing in a room with a wall paper giving the thumbs up |
|  | a man in a t - shirt and shorts is making a hand gesture | a man in a t - shirt and shorts is putting his finger on his lips |
|  | a man making a hand gesture in a room | a man making an ok sign with his hands in a room |
|  | a woman in a pink shirt making a hand gesture | a woman in a pink shirt giving an open palm with five fingers isolated |

Figure 2: Comparison of dataset before and after the post-processing of LLAMA3.

ditionally, some gestures involve interaction with other parts of the body, such as ‘mute,’ where the index finger must be placed on the lips while the rest of the hand forms a fist. Under these conditions, we input the same text prompt into Stable Diffusion (baseline) and compared the generated images with those from Hand1000. The results demonstrated that our approach significantly outperformed Stable Diffusion. A detailed analysis of each image will follow.

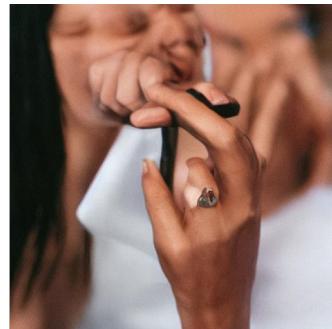
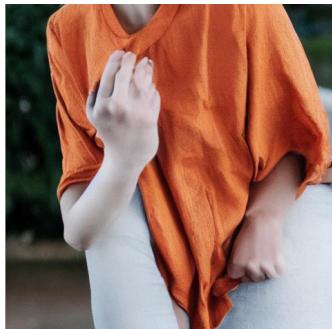
Figure 3 and 4 present gestures “call” and “four.” The main challenge of the “call” gesture lies in correctly depicting the extended thumb and little finger, with the remaining fingers clenched. Hand1000 successfully captures this feature, while Stable Diffusion struggles, producing distorted hands with abnormal finger numbers and shapes. Additionally, Hand1000 effectively conveys other elements of the text prompt, such as “girl,” “woman,” and “man,” accurately illustrating the heads and bodies of these individuals in a realistic manner. In contrast, Stable Diffusion only depicts facial features in the third image, with the first two showing no facial details. Regarding the clothing information, including “orange shirt,” “red shirt,” and “robe,” Hand1000 consistently delivers, whereas Stable Diffusion only correctly presents the “orange shirt” in the first image. The second

image contains disorganized, hand-like shapes, and in the third, the “man” appears to be draped in a sheet rather than a robe. For the “four” gesture, the challenge is to draw four extended fingers with the thumb naturally bent toward the palm. Hand1000 achieves this, while Stable Diffusion fails to reflect the gesture’s characteristics. Moreover, Hand1000 accurately portrays other aspects of the prompt, such as the “face mask” in the final image.

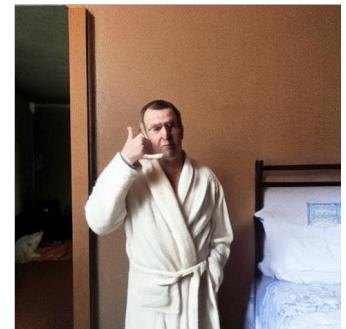
Figure 5 and 6 depict the “like” and “mute” hand gestures, both of which share a common challenge: while the emphasis is on extending a single finger, the true difficulty lies in rendering the remaining four fingers in a realistic and anatomically accurate manner. Hand1000 excels in this regard, as even when zoomed in, the images demonstrate perfect occlusion, curvature, and interaction between the fingers. In contrast, Stable Diffusion’s representation of the hand is notably poor. Furthermore, Hand1000 effectively captures other details from the text prompt, such as “a puzzle heart on her chest,” “in front of a truck,” “with a beard,” and “with glasses.” The “mute” gesture presents an additional challenge in requiring the index finger to be placed against the lips. Hand1000 handles the relationship between the finger and lips seamlessly, avoiding any unnatural fusion or distortion. On the other hand, images generated by Stable Diffusion either fail to position the finger on the lips or exhibit unrealistic merging of the finger and lips.

Figure 7 and 8 illustrate the “OK” and “palm” hand gestures. The “OK” gesture is relatively complex, requiring three fingers to be extended while the thumb and index finger form a circular shape. Hand1000 accurately represents this gesture without any distortion. Moreover, it effectively handles lighting, shadows, and occlusion in the hand area, with one image (second row, third column of Figure 7) showcasing impressive detail—even the joints and wrinkles of the hand are visible upon zooming in. The “palm” gesture, though seemingly simple, involves an open hand with five extended fingers, which requires precise modeling of finger count, length, and spatial relationships. For example, in the first row, second column of Figure 8, the image generated by Stable Diffusion displays realistic hand texture and color, but the finger count is incorrect, and the hand is distorted. In contrast, Hand1000 produces images with accurate finger count and natural hand characteristics, achieving excellent results.

Stable Diffusion



Hand1000

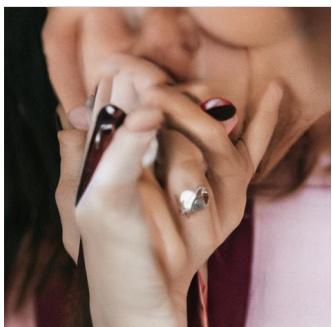


a girl in an orange shirt making phone call hand gesture

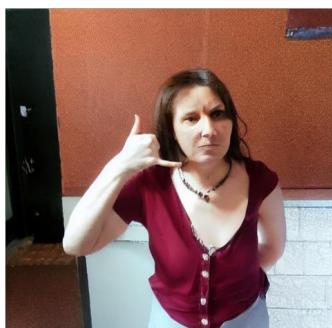
a woman in a red shirt making phone call hand gesture

a man in a robe standing in a bedroom making phone call hand gesture

Stable Diffusion



Hand1000



a woman in a maroon shirt making phone call hand gesture

a woman in a blue shirt making phone call hand gesture

a woman is making phone call hand gesture while standing in front of a purple wall

Figure 3: Comparison of images in hand gesture of "phone call" generated by stable diffusion and our Hand1000.

Stable Diffusion



Hand1000



a woman in a blue dress making four fingers up hand gesture

a man standing in front of a tv making four fingers up hand gesture

a woman wearing a face mask and making four fingers up hand gesture

Stable Diffusion



Hand1000



a woman in a green dress making four fingers up hand gesture

a woman in a white dress making four fingers up hand gesture

a girl in a white shirt standing in front of a bookshelf making four fingers up hand gesture

Figure 4: Comparison of images in hand gesture of "four" generated by stable diffusion and our Hand1000.

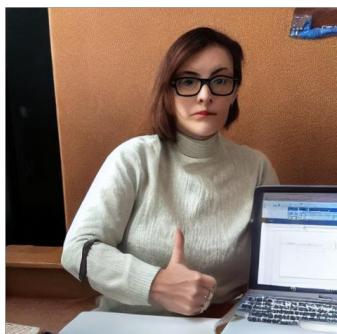
Stable Diffusion



Hand1000



Stable Diffusion



Hand1000

a woman in glasses and a sweater is sitting at a desk giving the thumbs up

a woman giving the thumbs up sign while sitting at a table

a man giving the thumbs up in front of a tree

Figure 5: Comparison of images in hand gesture of "like" generated by stable diffusion and our Hand1000.

Stable Diffusion



Hand1000



a man with a beard
sitting on a bed with
his finger on his lips

a girl with glasses
and a white shirt is
putting her finger on
her lips

a man with a beard
and a white shirt is
making a sign with
his finger on his lips

Stable Diffusion



Hand1000



a girl is holding her
finger to her mouth
with her finger on her
lips

a man with his finger
on his lips

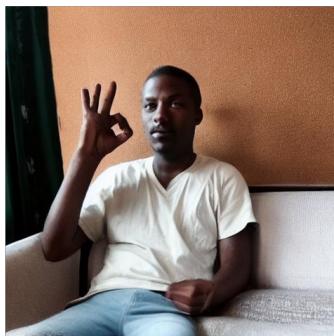
a young woman is
making a sign with
her finger on her lips

Figure 6: Comparison of images in hand gesture of "mute" generated by stable diffusion and our Hand1000.

Stable Diffusion



Hand1000



a man sitting on a couch with his hand up making an ok sign with his hands

a woman in a green sweater and grey skirt making the ok sign with her hands

a man in glasses making the ok sign with his hands

Stable Diffusion



Hand1000



a man in a striped shirt and blue shorts is making an ok sign with his hands

a woman in a brown sweatshirt making an ok sign with her hands

a man in a grey sweatshirt making an ok sign with his hands

Figure 7: Comparison of images in hand gesture of "ok" generated by stable diffusion and our Hand1000.

Stable Diffusion



Hand1000



a woman in a purple robe sitting on a bed giving an open palm with five fingers isolated

a woman in a white shirt is giving an open palm with five fingers isolated

a woman with long curly hair and a black top giving an open palm with five fingers isolated

Stable Diffusion



Hand1000



a woman in a green shirt giving an open palm with five fingers isolated

a woman in a blue shirt giving an open palm with five fingers isolated

a woman with red hair and a shirt giving an open palm with five fingers isolated

Figure 8: Comparison of images in hand gesture of "palm" generated by stable diffusion and our Hand1000.

References

- Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Mishra, S.; Seth, S.; Jain, S.; Pant, V.; Parikh, J.; Jain, R.; and Islam, S. M. 2024. Image Caption Generation using Vision Transformer and GPT Architecture. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 1–6. IEEE.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Open and efficient foundation language models. *Preprint at arXiv*. <https://doi.org/10.48550/arXiv.2302>.