

FoundHand: Large-Scale Domain-Specific Learning for Controllable Hand Image Generation

Supplementary Material



Figure 10. (a) [105] shows stochastic condition (SC) improves 3D consistency. (b) Encoding pose as skeleton images or numerical values provides less accurate control than our representation. (c) Naively training on more hand data still fails like most diffusion models. (d) reference condition is dominated by the pose condition and output fails to replicate the reference. (e) SD 3.5-Large still regularly fails on hands.

A. Experiment Details and Ablation

We train FoundHand on $4\times H100$ for 3 days with FP32 precision. At inference, we sample an image using 250 DDPM steps with CFG weight $w = 2.5$, which takes 20s on an RTX3090 or 10s using 100 DDIM steps.

In Figure 10, we show the ablation studies of our model. We found that without stochastic conditioning (SC), the video generation has degraded quality. Encoding 2D keypoints as 2D heatmaps and spatially aligning the image features, heatmaps, and mask improve fidelity to the pose control. Without applying condition dropout during training, we found that the pose condition would strongly dominate the reference condition, impairing the appearance control ability of FoundHand. We also tried simply scaling up the dataset and finetuning StableDiffusion UNet but the result is unsatisfying. Compared with our hand domain-specific model, even the latest SD 3.5-Large still often produces wrong hand anatomy.

B. Dataset Stats

FoundHand-10M comprises of 72.9% lab-captured, 10.2% synthetic, 16.9% in-the-wild images. 80.9% of our dataset has multiview video sequences and 53.3% has both ego and exocentric views. 64.9% of the scenes has hand-object interaction and 22.4% has hand-hand interaction. 81.6% has indoor scenes.

C. Gesture Transfer

In Fig. 11, we showcase results on Gesture Transfer application. FoundHand generates high quality gesture transferred images and faithfully follows the reference appearance and

the target hand pose. In contrast, the baselines which were trained on hands shows limited ability to preserve reference appearance and generates distorted fingers. Moreover, we try to quantitatively measure the generation ability of our models against baselines in Tab. 1. Therefore, we proposed the identity generation in the main paper, and we show the visual result in Fig. 12. We found that FoundHand achieves almost perfect identity generation in all non-cherry picked images, while our baselines [71, 94, 96] fails to generate high fidelity results in some times or all times.

D. Domain Transfer

In Fig. 13, we demonstrate more examples of domain transferred images from ReInterhand [55]’s challenging two-hand poses to EpicKitchen [13]’s appearances and backgrounds. Thanks to strong generalization and appearance preservation of FoundHand, our model can be a useful data augmentation tool for dexterity learning. Specifically, fine-tuned with our domain transferred images, the off-the-shelf hand estimation model [65] shows higher fidelity in very challenging hand images than before fine-tuned.

E. Novel View Synthesis (NVS)

We showcase more qualitative examples of novel view synthesis using our model on InterHand2.6M [54] (Fig. 14 and Fig. 15) and web-sourced in-the-wild images (Fig. 16). The motivation was that the existing general-purpose methods for novel view synthesis from single images shows very poor quality and fidelity in hands, because of the complex articulation of fingers. This avoids applications in understanding humans, AR/VR, and human-robot interactions. Therefore, we repurpose our model to provide realistic novel view synthesis of hands, given only a single image. Without being explicitly trained on any 3D representations or NVS data, our model can produce remarkably reliable NVS results, even with backgrounds and difficult hand poses. Notably, compared with baselines [84, 103] leveraging NeRF [51] to ensure 3D consistency, our geometry-free image generative model demonstrates great 3D understanding of hand. We found this showing robust 3D prior of hand without explicit 3D geometric context such as depth or mesh template.

F. HandFixer

In Fig. 18, we demonstrate more results in fixing malformed hands. Generative models could produce malformed hands

such as non-five fingers and distorted hand structures. Compared with task-specific methods like HandRefiner [47] and RealisHuman [101] which requires accurate 3D hand estimation, our FoundHand performs zero-shot hand fixing, demonstrating exceptional generalization to diverse artistic and abstract styles. FoundHand can even work with sketches of hands, drawings, challenging hand appearances, and difficult poses interacting with objects. The model also shows better understanding of the context, particularly preserving the hand-object interaction context after fixing the hand. Moreover, FoundHand only requires masks where users want to changes and 2D hand keypoints, which is different from [47, 101] who asks 3D hand models. This enables more flexible and easier controls for the users to fix hands.

G. Hand Video and Hand-Object Interaction

Given the first frame image and a sequence of 2D keypoints captured in the wild by an iphone camera, FoundHand can autoregressively generate a motion-controlled video, despite not explicitly trained on videos. This shows our model’s high versatility and potentials for being used in various applications. We provide hand video synthesis results in more details in Fig. 19, where ControlNeXt [67] and AnimateAnyone [31] struggles to follow the pose change or present significant visual artifacts, while our model demonstrate robust generalization and emergent understanding of some physical effects such as casted shadows.

Fig. 17 compares our models’ ability to generate hand-object interaction (HOI) videos again the state-of-the-art HOI video synthesis model. Note that this involves object translation and deformable objects. FoundHand has naturally seen many hand-object interaction and manipulation scenes and surprisingly develops emergent physical understanding of HOI (object translation and deformation.) without explicit knowledge of the object context. On the other hand, CosHand [94] is trained on a set of specific data consisting of before-after pairs of HOI focusing on interaction-induced change. However, it shows some overfitting such as random objects which we guess were from their training distribution.

Please see supplementary video for video results.

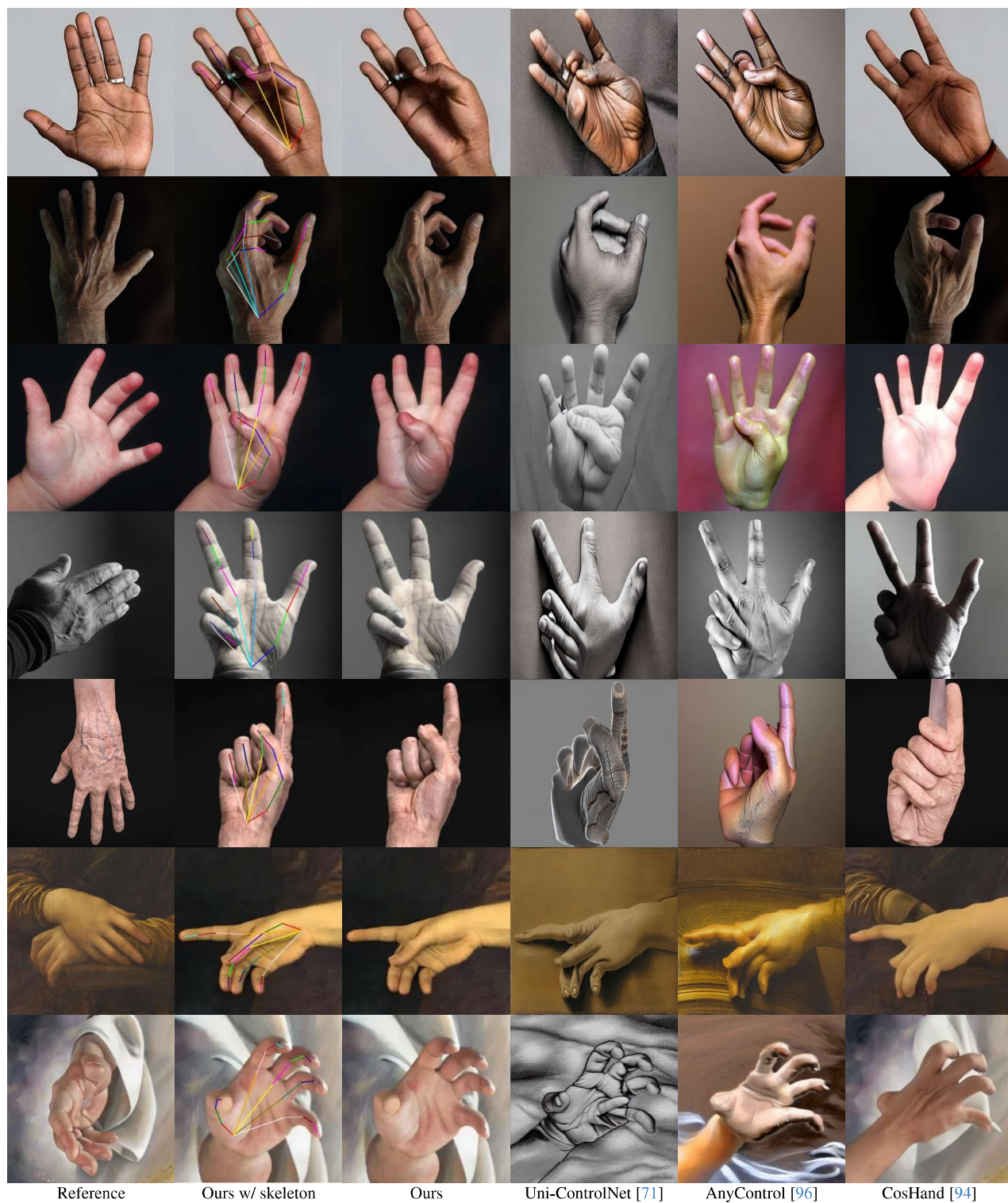


Figure 11. More results on Gesture Transfer application. FoundHand generates high quality gesture transferred images and faithfully follows the reference appearance and the target hand pose. In contrast, the baselines which were trained on hands shows limited ability to preserve reference appearance and generates distorted fingers.



Figure 12. Identity generation to quantitatively measure the generation ability of FoundHand and baselines. FoundHand achieves almost perfect reconstruction in all examples, while the baselines occasionally or always fail to generate high fidelity results.



Figure 13. FoundHand can provide domain transfer from highly complex hand poses from synthetic data [55] and reference images from real-world data [13]. Fine-tuned with our domain transferred images, the off-the-shelf hand estimation model [65] shows even higher fidelity in very challenging hand images.



Figure 14. We test novel view synthesis (NVS) on the test data split of InterHand2.6M. Compared with baselines [84, 103] leveraging NeRF [51] to ensure 3D consistency, our geometry-free image generative model demonstrates great 3D understanding of hand.

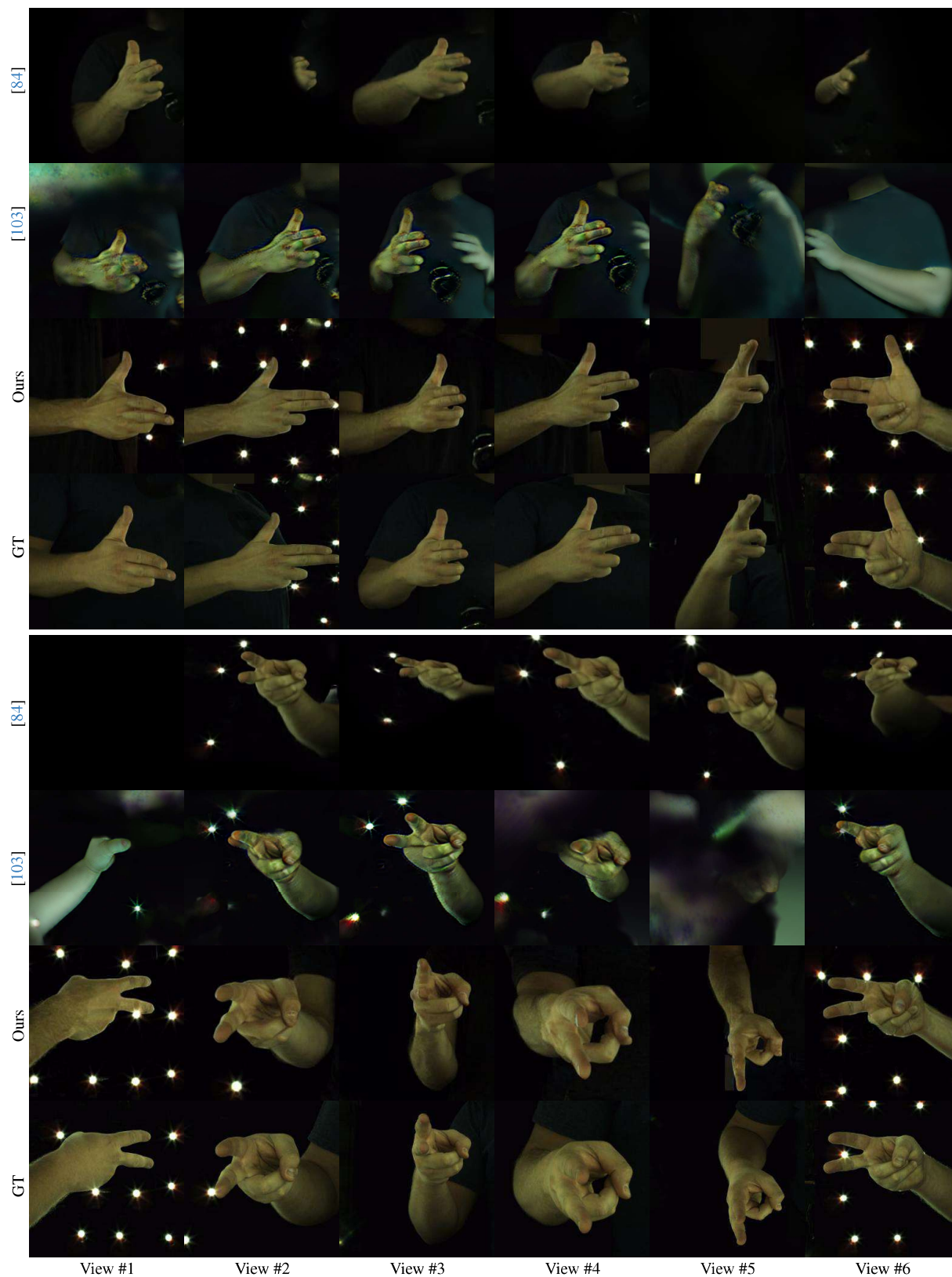


Figure 15. Fig. 14 continued.



Figure 16. Internet-sourced single image to Novel View Synthesis. FoundHand can provide reasonable novel view synthesis, showing robust 3D prior of hand without explicit 3D geometric context such as depth or mesh template.

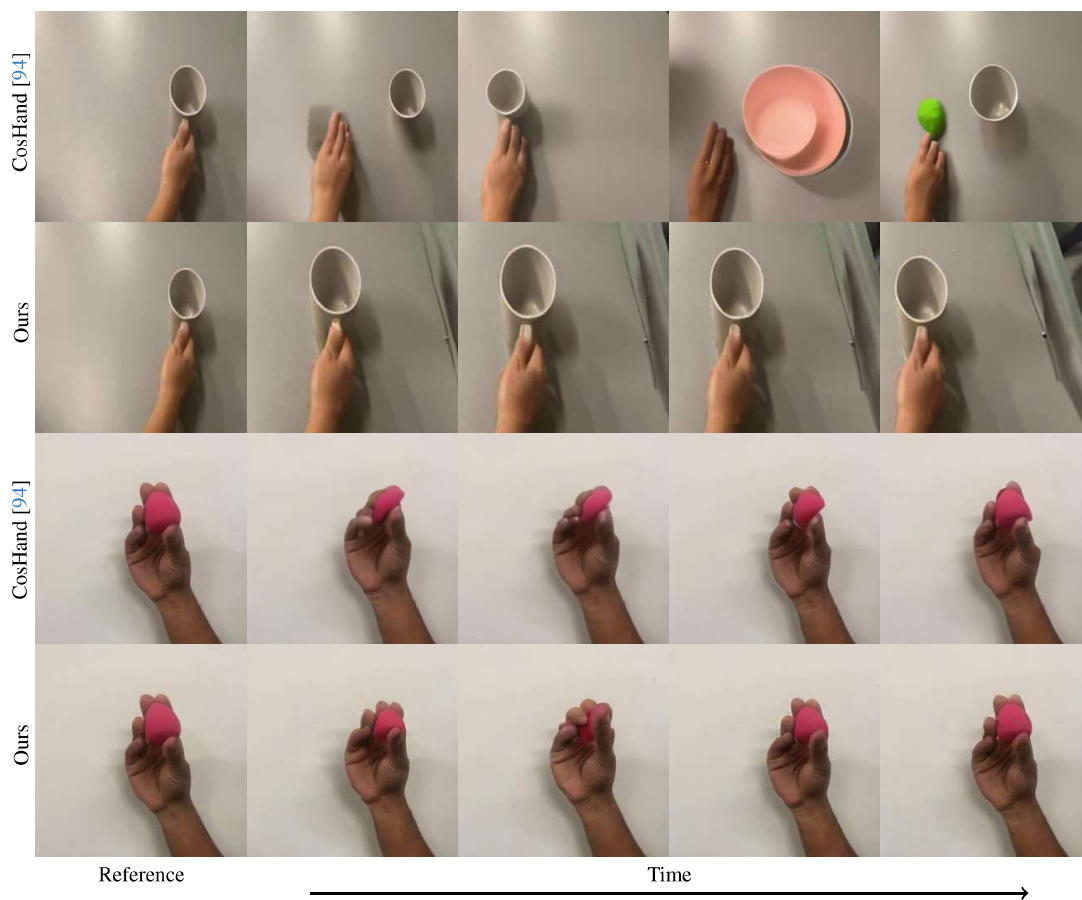


Figure 17. FoundHand has naturally seen many hand-object interaction and manipulation scenes and surprisingly develops emergent physical understanding of HOI (object translation and deformation.) without explicit knowledge of the object context. On the other hand, CosHand [94] is trained on specific HOI data focusing on interaction-induced change but shows some overfitting (1st row).

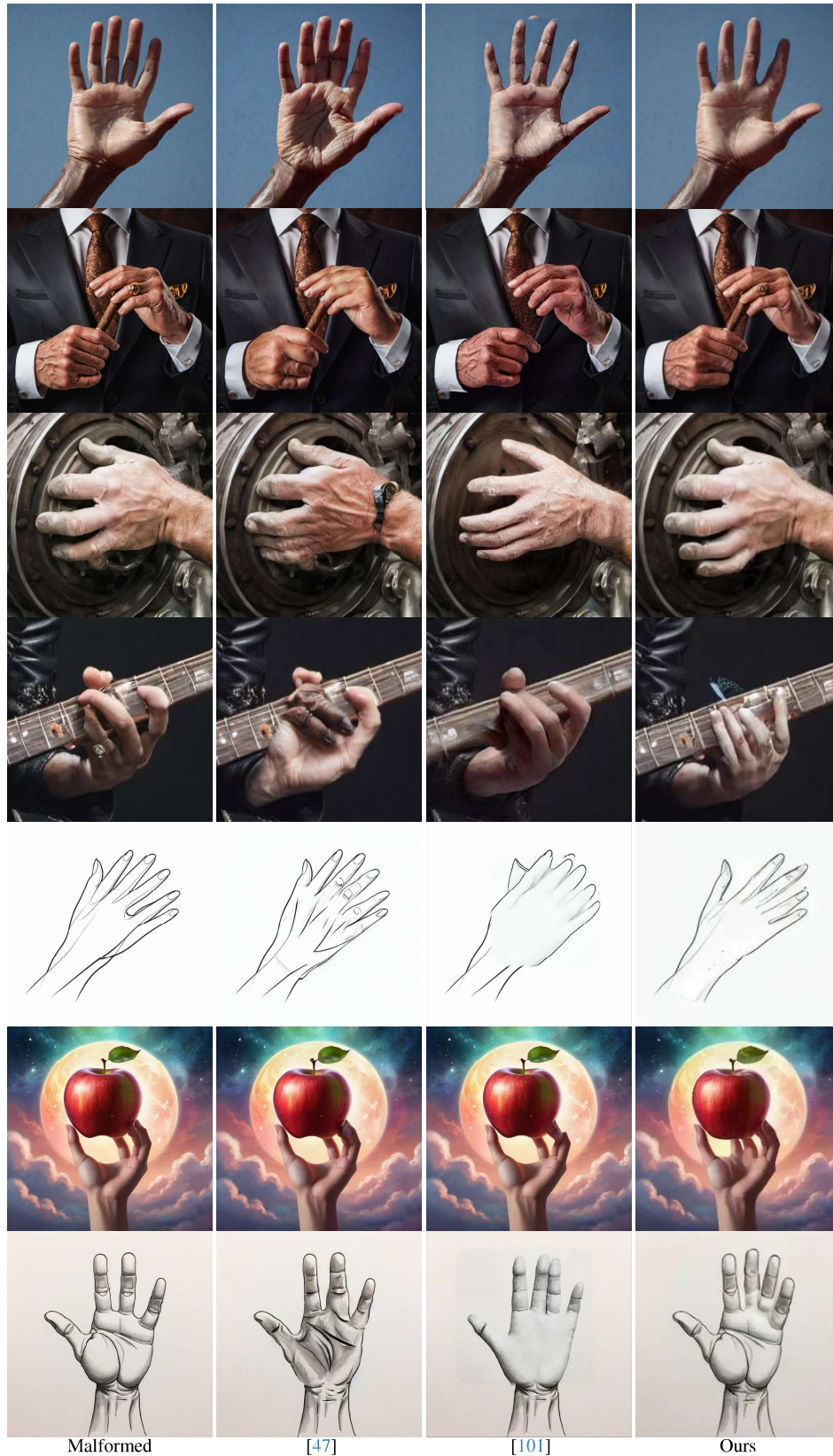


Figure 18. Compared with task-specific methods like HandRefiner [47] and RealisHuman [101] which requires accurate 3D hand estimation, our FoundHand performs zero-shot hand fixing, demonstrating exceptional generalization to diverse artistic and abstract styles (5th and 7th row). Our model also shows better understanding of the context, particularly preserving the hand-object interaction context after fixing the hand (2nd, 4th, and 6th row).

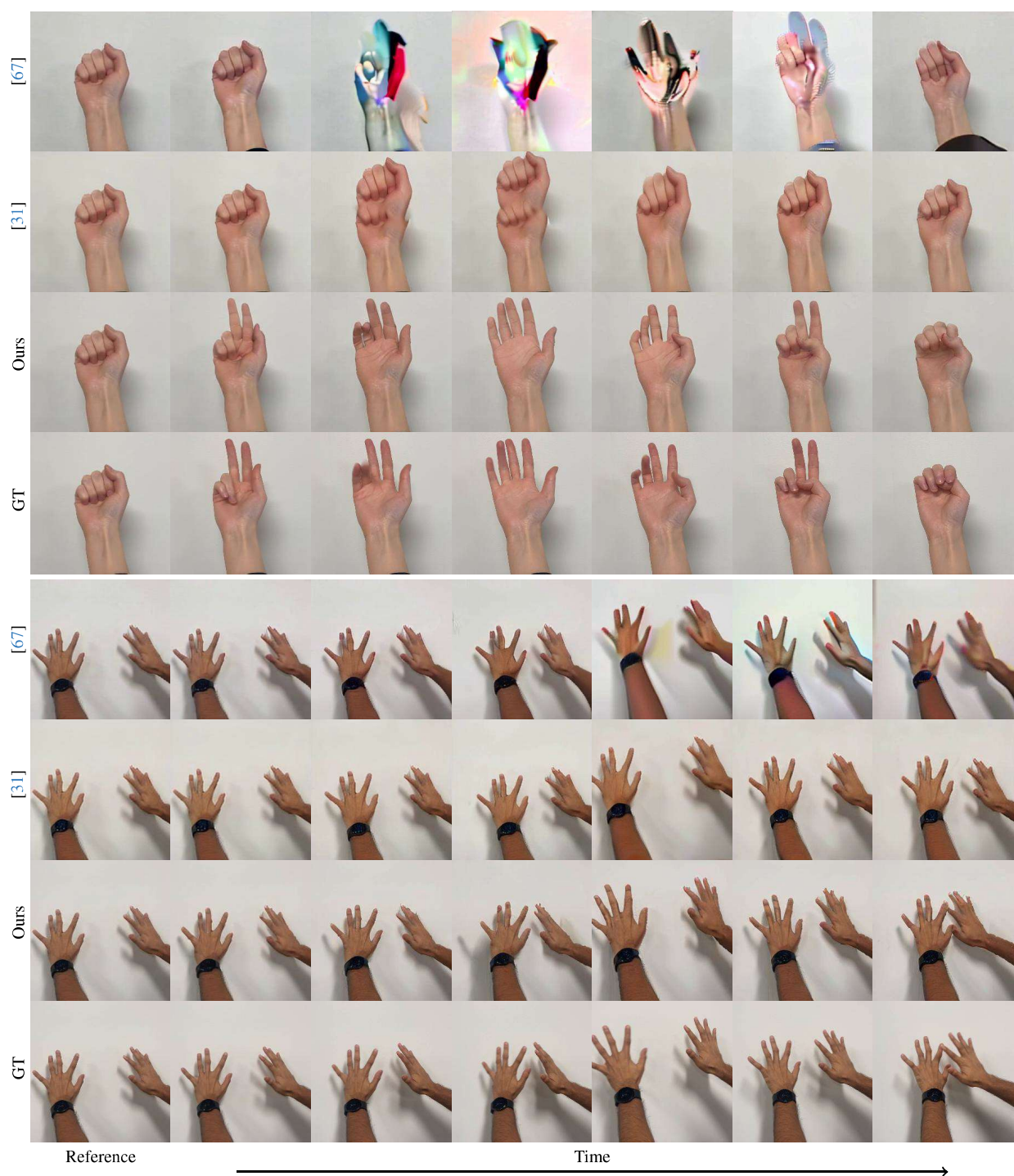


Figure 19. Given the first frame image and a sequence of 2D keypoints captured in the wild by an iphone camera, FoundHand can autoregressively generate a motion-controlled video, despite not explicitly trained on videos. This shows our model’s high versatility and potentials for being used in various applications. ControlNeXt [67] and AnimateAnyone [31] struggles to follow the pose change or present significant visual artifacts while our model demonstrate robust generalization and emergent understanding of some physical effects such as casted shadows.