1. Supplementary Material

A. Overview

The supplementary material is organized into the following sections:

- Section B: Implementation Details
- Section C: Ablation for VQ-MANO Pose Tokenizer
- Section D: In-the-Wild Reconstruction Evaluation
- Section E: Occluded and Masked Hands Reconstruction
- Section H: Effectiveness of Proposed MaskHand Components
- Section F: Impact of 2D Pose Context
- Section G: Confidence-Aware Unconditional Mesh Generation
- Section I: Masking Ratio during Training
- Section J: Effectiveness of Expectation-Approximated Differential Sampling
- Section K: Confidence-Guided Masking
- Section L: Impact of VQ-MANO Tokenizer on MaskHand
- Section M: Impact of Multi-Scale Features
- Section N: Deformable Cross-Attention Layers in MaskHand
- Section O: Qualitative Results in the Wild

Project website can be founds at https://m-usamasaleem.github.io/publication/MaskHand/MaskHand.html.

B. Implementation Details

The implementation of MaskHand, developed using PyTorch, comprises two essential training phases: the VQ-MANO tokenizer and the context-guided masked transformer. These phases are meticulously designed to ensure accurate 3D hand mesh reconstruction while balancing computational efficiency and model robustness.

VQ-MANO. In the first phase, the VQ-MANO module is trained to learn discrete latent representations of hand poses. The pose parameters, $\theta \in \mathbb{R}^{16\times3}$, encapsulate the global orientation ($\theta_1 \in \mathbb{R}^3$) and local rotations ($[\theta_2, \ldots, \theta_{16}] \in \mathbb{R}^{16\times3}$) of hand joints. The architecture of the tokenizer employs ResBlocks [5] and 1D convolutional layers for the encoder and decoder, with a single quantization layer mapping continuous embeddings into a discrete latent space. To train the hand pose tokenizer, we utilized a range of datasets capturing diverse hand poses, interactions, and settings. Specifically, we leveraged DexYCB [3], InterHand2.6M [9], MTC [14], and RHD [15]. These datasets collectively provide a rich spectrum of annotated data, enabling the model to generalize effectively across various real-world scenarios. The training process spans 400K iterations and uses the Adam optimizer with a batch size of 512 and a learning rate of 1×10^{-4}. The loss function combines reconstruction and regularization objectives, weighted as $\lambda_{\text{recon}} = 1.0$, $\lambda_{\text{E}} = 0.02$, $\lambda_{\theta} = 1.0$, $\lambda_{\text{V}} = 0.5$, and $\lambda_{\text{J}} = 0.3$. The final pose tokenizer is trained on DexYCB, InterHand2.6M, MTC, and RHD datasets, resulting in a model with 64 tokens and a codebook size of 2048×256.

Context-Guided Masked Transformer. The second phase involves training the context-guided masked transformer, with the pose tokenizer frozen to leverage its pre-trained pose priors. This phase is dedicated to synthesizing pose tokens conditioned on input images and refining the 3D mesh reconstruction. Multi-resolution feature maps at $1\times$ and $4\times$ scales are used to capture both global and local contextual details, allowing the model to handle complex hand articulations and occlusions. The overall architecture of the system, including the Graph-based Anatomical Pose Refinement (GAPR) and Context-Infused Masked Synthesizer. The GAPR consists of two blocks (xB) of graph transformers to effectively model joint dependencies and ensure anatomical consistency. Meanwhile, the Context-Infused Masked Synthesizer employs four transformer layers (xN) to integrate multi-scale image features and refine pose token predictions through deformable cross-attention and token dependencies. The default number of iterations in Confidence-Guided Sampling is 5, which we use for the ablation study. The overall loss function integrates multiple objectives to guide the model toward robust reconstructions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{MANO}} + \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}},$$

where $\mathcal{L}_{\text{mask}}$ minimizes errors in masked token predictions, $\mathcal{L}_{\text{MANO}}$ ensures consistency in MANO shape (β) and pose (θ) parameters, $\mathcal{L}_{\text{3D}}$ aligns the predicted and ground-truth 3D joint positions, and $\mathcal{L}_{\text{2D}}$ preserves accurate 2D joint projections. The loss weights are configured as $\lambda_{\text{mask}} = 1.0$, $\lambda_{\text{MANO}} = 1.5 \times 10^{-3}$ (with $\lambda_{\theta} = 1 \times 10^{-3}$ for pose and $\lambda_{\beta} = 5 \times 10^{-4}$

for shape), $\lambda_{3D} = 5 \times 10^{-2}$, and $\lambda_{2D} = 1 \times 10^{-2}$. This phase is trained for 200K iterations using the Adam optimizer on NVIDIA RTX A6000 GPUs with a batch size of 48 and a learning rate of 1×10^{-5}.

Generalization to Text-to-Mesh Generation Details. The MaskHand model is designed to be modular and adaptable, extending beyond image-conditioned tasks to support text-to-mesh generation. To achieve this, we replaced image-based conditioning with text guidance, enabling MaskHand's Masked Synthesizer to generate diverse 3D meshes directly from textual input. Additionally, we explored the model's capability to synthesize meshes without 2D pose guidance, making it rely solely on text prompts for generation. For training, we used the American Sign Language (ASL) dataset, a widely recognized resource for hand-based sign language recognition in English-speaking regions such as the United States and Canada. The dataset consists of 26 one-handed gestures representing the alphabet, making it suitable for text-to-mesh experiments. Specifically, we used the ASL alphabet dataset from Kaggle [1] for training. Since the ASL dataset lacks 3D annotations (e.g., MANO parameters), we leveraged MaskHand to generate pseudo-ground-truth (p-GT) annotations, which were then used to train the text-guided version of the model. To integrate textual information, we extracted CLIP [12] embeddings from ASL labels, enabling seamless text-based conditioning within the generative pipeline. During testing, we applied a top-5% probabilistic sampling strategy, allowing the model to generate multiple plausible meshes per text input while ensuring diversity and consistency in synthesis.

B.1. Data Augmentation

In the initial training phase, the VQ-MANO module leverages prior knowledge of valid hand poses, serving as a critical foundation for the robust performance of the overall MaskHand pipeline. To deepen the model's understanding of pose parameters, hand poses are systematically rotated across diverse angles, enabling it to effectively learn under varying orientations. In the subsequent training phase, the robustness of MaskHand is further enhanced through an extensive augmentation strategy applied to both input images and hand poses. These augmentations—such as scaling, rotations, random horizontal flips, and color jittering—introduce significant variability into the training data. By simulating real-world challenges like occlusions and incomplete pose information, these transformations prepare the model for complex, unpredictable scenarios. This comprehensive approach to data augmentation is a cornerstone of the training process, significantly improving the model's ability to generalize and produce reliable, precise 3D hand mesh reconstructions across a wide range of conditions.

B.2. Camera Model

In the MaskHand pipeline, a simplified perspective camera model is employed to project 3D joints onto 2D coordinates, striking a balance between computational efficiency and accuracy. The camera parameters, collectively represented by Π, include a fixed focal length, an intrinsic matrix $K \in \mathbb{R}^{3\times3}$, and a translation vector $T \in \mathbb{R}^3$. To streamline computations, the rotation matrix R is replaced with the identity matrix I_3, further simplifying the model. The projection of 3D joints J_{3D} onto 2D coordinates J_{2D} is described as $J_{2D} = \Pi(J_{3D})$, where the operation encapsulates both the intrinsic parameters and the translation vector. This modeling approach reduces the parameter space, enabling computational efficiency while maintaining the accuracy required for robust 3D hand mesh reconstruction. By focusing on the most critical components, the model minimizes complexity without compromising performance.

C. Ablation for VQ-MANO

Tables 1 and 2 summarize an ablation study on the Freihand [16] dataset, focusing on two key parameters: the number of pose tokens and the codebook size. Table 1 shows that increasing the number of pose tokens, while fixing the codebook size at 2048×256, improves performance significantly, reducing PA-MPJPE from 1.01 mm to 0.41 mm and PA-MPVPE from 0.97 mm to 0.41 mm as tokens increase from 16 to 128. Table 2 highlights the effect of increasing the codebook size with a fixed token count of 64, showing a reduction in PA-MPJPE from 0.66 mm to 0.43 mm and PA-MPVPE from 0.65 mm to 0.44 mm as the size grows from 1024×256 to 4096×256. Notably, the codebook size has a stronger impact on performance than the number of pose tokens. The final configuration, with a codebook size of 2048×256 and 64 tokens, balances efficiency and accuracy, achieving PA-MPJPE of 0.47 mm and PA-MPVPE of 0.44 mm. These results emphasize the importance of jointly optimizing these parameters for effective hand pose tokenization.

D. In-the-Wild Reconstruction Evaluation

Table 3 presents a zero-shot evaluation comparing MaskHand with recent state-of-the-art methods on the challenging HInt benchmark using the PCK metric. MaskHand consistently achieves the best results across all subsets—NewDays, VISOR, and Ego4D—and evaluation criteria (All Joints and Visible Joints). Notably, MaskHand significantly surpasses HaMeR

Table 1. Stage-1: Impact of Number of Pose Tokens (Codebook = 2048 × 256) on VQ-MANO on Freihand dataset

	Number of Pose Tokens			
Metric	**16**	**32**	**64**	**128**
PA-MPJPE (mm)	1.01	0.59	0.47	0.41
PA-MPVPE (mm)	0.97	0.57	0.44	0.41

Table 2. Stage-1: Impact of Number of Codebook Size (Pose Tokens = 64) on VQ-MANO on Freihand dataset

	Number of Codebook Size			
Metric	**1024 × 256**	**2048 × 128**	**2048 × 256**	**4096 × 256**
PA-MPJPE (mm)	0.66	0.56	0.47	0.43
PA-MPVPE (mm)	0.65	0.58	0.44	0.44

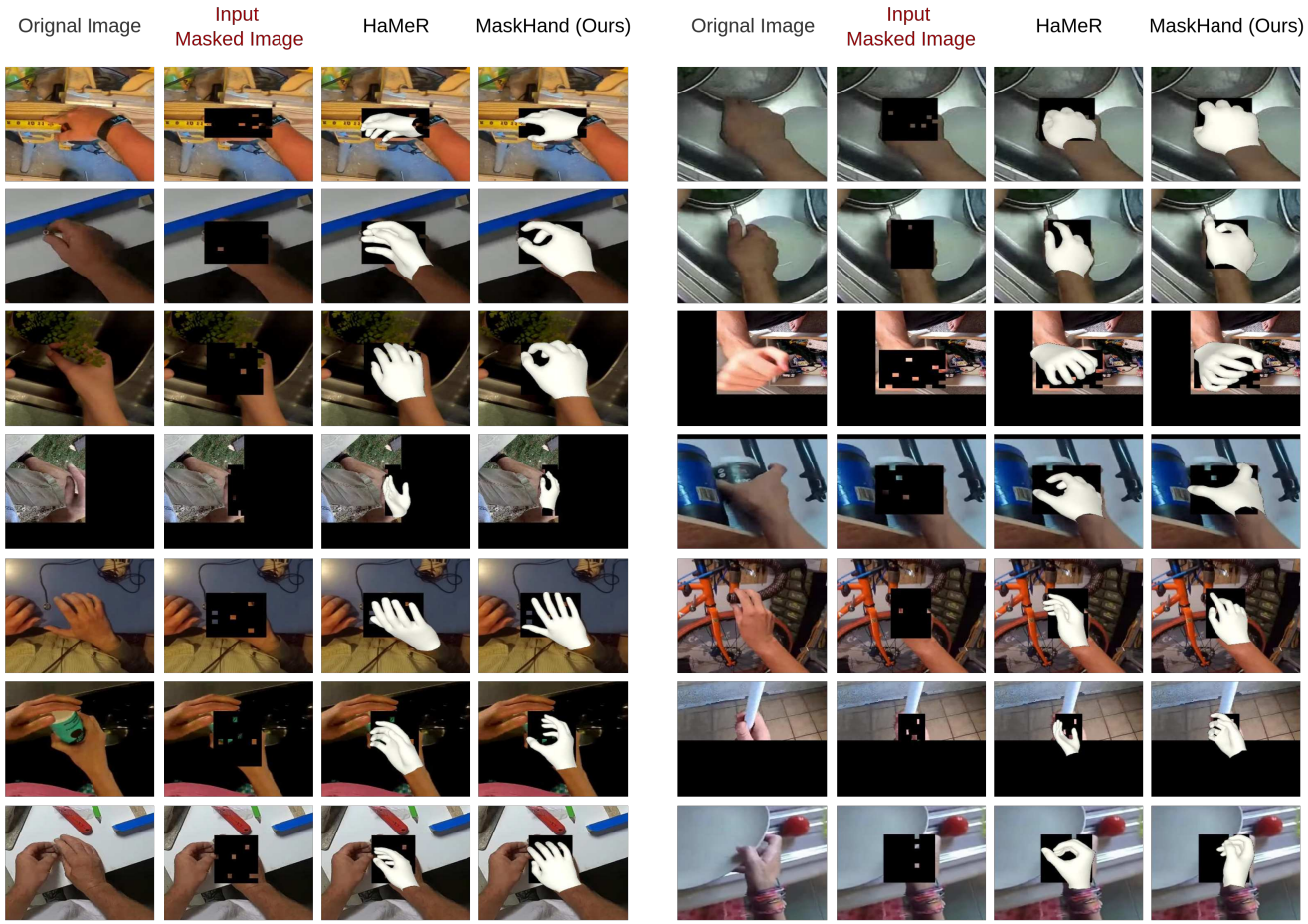

Figure 1. SOTA Comparison: Qualitative zero-shot evaluation on HInt Benchmark [11] for heavily masked hand images. MaskHand reconstructs occluded hand poses, demonstrating robustness to severe occlusions and generalizability to unseen masked regions.

(previous SOTA), demonstrating improvements of up to 7.5% at the strictest threshold (PCK@0.05) on Ego4D (46.4% vs. HaMeR's 38.9%) and 3.1% on VISOR (46.1% vs. 43.0%) for All Joints. Similar trends appear with Visible Joints, where MaskHand improves by 5.6% on VISOR (62.1% vs. HaMeR's 56.6%) and 7.3% on Ego4D (59.3% vs. 52.0%). These substantial gains underscore MaskHand's superior accuracy and robustness in reconstructing hands under challenging, real-world occlusions. While the model demonstrates impressive performance, the results also highlight areas for future improvements,

particularly in severely occluded regions, where further advancements could provide additional gains in accuracy.

Table 3. Zero-Shot In-the-Wild Robustness Evaluation on the HInt Benchmark [11] using the PCK Metric: Comparison with SOTA Methods. **None of the models were trained on or have previously seen the HInt dataset.**

Method	Venue	NewDays			VISOR			Ego4D		
		@0.05 (↑)	@0.1 (↑)	@0.15 (↑)	@0.05 (↑)	@0.1 (↑)	@0.15 (↑)	@0.05 (↑)	@0.1 (↑)	@0.15 (↑)
All Joints										
FrankMocap [13]	*ICCVW 2021*	16.1	41.4	60.2	16.8	45.6	66.2	13.1	36.9	55.8
METRO [7]	*CVPR 2021*	14.7	38.8	57.3	16.8	45.4	65.7	13.2	35.7	54.3
MeshGraphormer [8]	*ICCV 2021*	16.8	42.0	59.7	19.1	48.5	67.4	14.6	38.2	56.0
HandOccNet (param) [10]	*CVPR 2022*	9.1	28.4	47.8	8.1	27.7	49.3	7.7	26.5	47.7
HandOccNet (no param) [10]	*CVPR 2022*	13.7	39.1	59.3	12.4	38.7	61.8	10.9	35.1	58.9
HaMeR [11]	*CVPR 2024*	48.0	78.0	88.8	43.0	76.9	89.3	38.9	71.3	84.4
MaskHand	**Ours**	**48.7**	**79.2**	**90.0**	**46.1**	**81.4**	**92.1**	**46.4**	**77.5**	**90.1**
Visible Joints										
FrankMocap [13]	*ICCVW 2021*	20.1	49.2	67.6	20.4	52.3	71.6	16.3	43.2	62.0
METRO [7]	*CVPR 2021*	19.2	47.6	66.0	19.7	51.9	72.0	15.8	41.7	60.3
MeshGraphormer [8]	*ICCV 2021*	22.3	51.6	68.8	23.6	56.4	74.7	18.4	45.6	63.2
HandOccNet (param) [10]	*CVPR 2022*	10.2	31.4	51.2	8.5	27.9	49.8	7.3	26.1	48.0
HandOccNet (no param) [10]	*CVPR 2022*	15.7	43.4	64.0	13.1	39.9	63.2	11.2	36.2	56.0
HaMeR [11]	*CVPR 2024*	60.8	87.9	94.4	56.6	88.0	94.7	52.0	83.2	91.3
MaskHand	**Ours**	**61.0**	**87.1**	**94.8**	**62.1**	**90.2**	**95.0**	**59.3**	**88.3**	**94.4**

E. Occluded and Masked Hands Reconstruction

Figure 1 provides a qualitative, zero-shot comparison between MaskHand and the state-of-the-art HaMeR method on severely masked images (with hand regions masked around 90%). Despite significant occlusion, MaskHand consistently generates more plausible and anatomically accurate hand reconstructions than HaMeR. Specifically, MaskHand effectively synthesizes occluded regions, demonstrating robust generalization to previously unseen masked areas and preserving natural hand poses and orientations. In contrast, HaMeR exhibits noticeable reconstruction failures, inaccuracies, and unnatural poses, particularly under extreme masking conditions. These results highlight MaskHand's superior capability in modeling uncertainty and synthesizing realistic meshes under severe occlusion, underscoring its robustness and practical effectiveness in challenging real-world scenarios.

F. Impact of 2D Pose Context

We investigate how the accuracy of the 2D pose estimator influences MaskHand's 3D reconstruction performance. In our main experiments, we utilize a lightweight OpenPose estimator OpenPose due to its computational efficiency and suitability for real-time applications. To quantify how improvements in 2D pose estimation may affect overall reconstruction quality, we conduct an additional analysis comparing OpenPose predictions against ground-truth 2D keypoints on the FreiHAND dataset. As shown in Table 4, using ground-truth 2D poses consistently yields better reconstruction metrics, notably improving both PA-MPJPE and PA-MPVPE. Although MaskHand achieves robust performance even with estimated keypoints, this analysis indicates that further advances in 2D pose estimation accuracy can directly enhance the quality of reconstructed 3D hand meshes.

Estimator	PA-MPJPE	PA-MPVPE	F@5mm	F@15mm
Ground Truth	5.2	5.1	0.834	0.993
2D OpenPose [2] Estimator	5.5	5.4	0.801	0.991

Table 4. Impact of 2D Pose Estimator on 3D Reconstruction Quality on Friehand dataset [16]

G. Confidence-Aware Unconditional Mesh Generation

MaskHand enables confidence-aware unconditional 3D hand mesh generation by leveraging generative masked modeling and probabilistic sampling. We generate 2,000 hand meshes by setting the image condition to zero in the image encoder, ensuring

Methods	APD(mm)↑	SI(%)↓
PCA	16.3	0.32
MaskHand (ours)	**19.3**	**0.04**

Table 5. Confidence-Aware Unconditional Mesh Generation

that synthesis is driven entirely by the learned pose distribution. Using Top-100 sampling, MaskHand not only produces high-quality, diverse meshes but also quantifies the confidence of each generated hand configuration. This confidence estimation allows MaskHand to distinguish between physically plausible and invalid meshes, an advantage over diffusion-based HHMR [6], which lacks direct plausibility quantification.

In contrast, PCA-based generation, which samples from the MANO parameter space, produces structurally valid but limited and less diverse hand poses due to its restriction to the linear PCA subspace. The qualitative results (Figure 2) highlight MaskHand's ability to synthesize highly articulated hand poses, while lower-confidence samples exhibit unnatural deformations, demonstrating its uncertainty quantification capability. Quantitatively, MaskHand achieves greater diversity than PCA, as reflected in a higher APD (19.3mm vs. 16.3mm), and generates more realistic meshes, reducing self-intersection (SI) from 0.32% to 0.04% (Table 5). These results confirm that MaskHand surpasses PCA in both diversity and realism, offering a principled approach to filtering implausible generations, making it a more reliable solution for unconditional 3D hand synthesis.

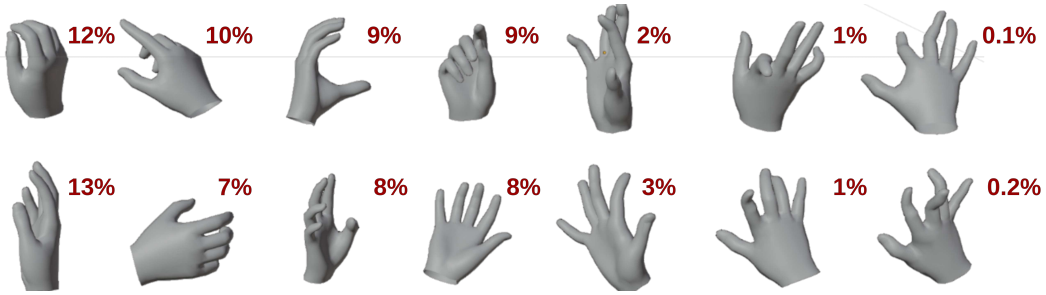

Figure 2. Confidence-Aware Unconditional Mesh Generation

H. Effectiveness of Proposed MaskHand Components

The ablation study on HO3Dv3 (Table 6) highlights GAPR as the most critical component, ensuring joint dependencies and anatomical coherence. Removing the Upsampler and 2D Pose Context slightly reduces accuracy, affecting fine details and spatial cues. The full MaskHand model achieves the best results, demonstrating the importance of these components for high-precision 3D hand mesh recovery. Qualitative comparisons in Figure 3 further illustrate these effects, showing the impact of each component on reconstruction quality.

Method	PA-MPJPE	PA-MPVPE	F@5mm	AUC_J	AUC_V
w/o. Upsampler	7.2	7.2	0.654	0.857	0.857
w/o. 2D Pose Context	7.1	7.1	0.656	0.857	0.858
w/o. GAPR	7.3	7.3	0.645	0.853	0.854
MaskHand (Full)	**7.0**	**7.0**	**0.663**	**0.860**	**0.860**

Table 6. Ablation study of testing results on the HO3Dv3 dataset [4] to evaluate the impact of proposed components. 'w/o' denotes 'without'.

I. Masking Ratio during Training

The ablation study in Table 7 shows that a broader masking range $\gamma(\tau \in \mathcal{U}(0, 0.7))$ achieves optimal results on HO3Dv3 and FreiHAND, with the lowest PA-MPVPE values of 7.0 and 5.5, respectively. This cosine-based masking strategy, where

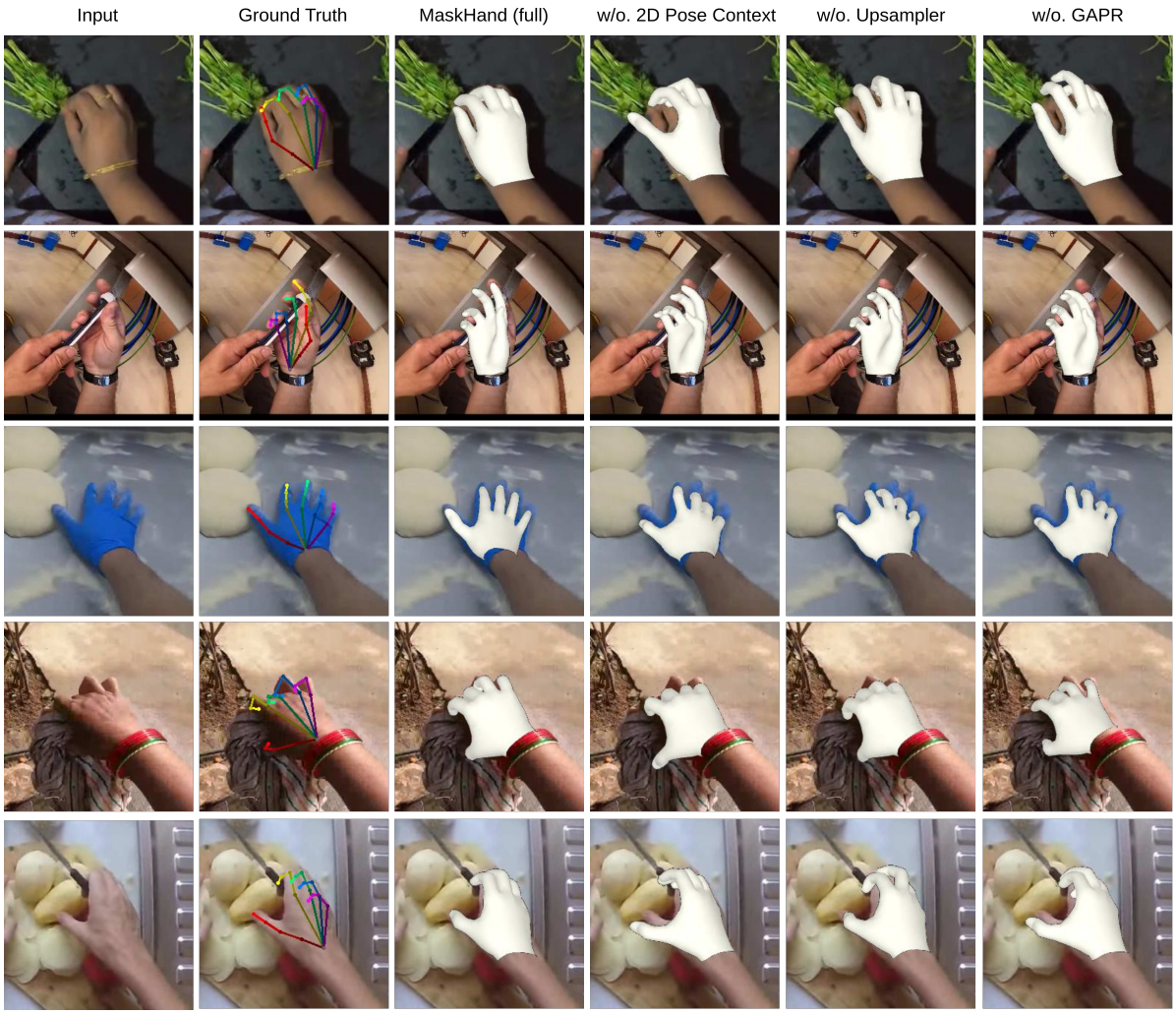

Figure 3. Qualitative ablation study on component impact: Full model achieves highest accuracy, validating each component's role.

the model learns to reconstruct from partially masked sequences, enhances robustness in 3D hand reconstruction. Narrower masking ranges, such as $\gamma(\tau \in \mathcal{U}(0, 0.3))$, increase error, highlighting the importance of challenging the model with broader masking for better generalization.

Table 7. Impact of masking ratio during training on HO3Dv3 [4] and FreiHAND [16] datasets.

Masking Ratio $\gamma(\tau)$	HO3Dv3		FreiHAND	
	PA-MPJPE	PA-MPVPE	PA-MPJPE	PA-MPVPE
$\gamma(\tau \in \mathcal{U}(0, 0.3))$	7.2	7.2	5.7	5.8
$\gamma(\tau \in \mathcal{U}(0, 0.5))$	7.1	7.1	5.6	5.6
$\gamma(\tau \in \mathcal{U}(0, 0.7))$	7.0	7.0	5.5	5.4
$\gamma(\tau \in \mathcal{U}(0, 1.0))$	7.2	7.3	5.8	5.7

J. Effectiveness of Expectation-Approximated Differential Sampling

The results presented in Table 8 highlight the critical role of Expectation-Approximated Differential Sampling in enabling accurate and robust 3D hand mesh recovery. The configuration utilizing all loss components—L_{mask}, L_{MANO}, L_{3D}, L_{2D},

and β—achieves the lowest PA-MPJPE and PA-MPVPE values of 0.70 mm on HO3Dv2 and 5.5 mm on FreiHAND, underscoring the importance of a holistic training approach. This configuration demonstrates the complementary strengths of L_{3D} in enforcing anatomical coherence, L_{2D} in mitigating monocular depth ambiguities, and L_{mask} in iteratively refining pose token predictions. Excluding critical components such as L_{3D} or L_{2D} leads to substantial degradation in performance, with errors rising to 8.1 mm on HO3Dv2 and 7.0 mm on FreiHAND. These results emphasize the necessity of these constraints for accurate 2D-to-3D alignment and plausible pose synthesis. Expectation-Approximated Differential Sampling is instrumental in this process, as it facilitates seamless integration of these losses by leveraging a differentiable framework for token refinement. This approach ensures that the latent pose space is effectively optimized, enabling the model to balance fine-grained token accuracy with global pose coherence. These findings validate the pivotal role of differential sampling in guiding the learning process, resulting in precise and confident 3D reconstructions under challenging scenarios.

Table 8. Impact of different loss combinations on PA-MPJPE and PA-MPVPE errors (in mm) for the HO3Dv2 and FreiHAND datasets.

Used Losses	**HO3Dv2**		**FreiHAND**	
	PA-MPJPE ↓	PA-MPVPE ↓	PA-MPJPE ↓	PA-MPVPE ↓
L_{mask}, β	7.6	7.5	6.1	6.3
$L_{mask}, L_{MANO}, \beta$	7.5	7.5	6.0	6.2
$L_{mask}, L_{MANO}, L_{3D}, L_{2D}, \beta$	7.0	7.0	5.5	5.4
L_{3D}, β	8.1	7.9	6.5	6.7
$L_{mask}, L_{MANO}, L_{2D}, \beta$	7.4	7.3	6.3	6.4

K. Confidence-Guided Masking

Figure 4 illustrates the iterative process of Confidence-Guided Sampling used during inference for refining pose predictions. The gray bars represent the total number of masked tokens across iterations, while the green line tracks the average confidence in the model's predictions. At the initial iteration, the majority of pose tokens remain masked, indicating high uncertainty. As iterations progress, the number of masked tokens decreases significantly, which aligns with a steady increase in the model's confidence. By the final iteration, only a minimal number of tokens remain masked, while the average confidence approaches its peak. This visualization highlights the systematic reduction in uncertainty and refinement of predictions over multiple iterations, enabling robust 3D pose reconstruction.

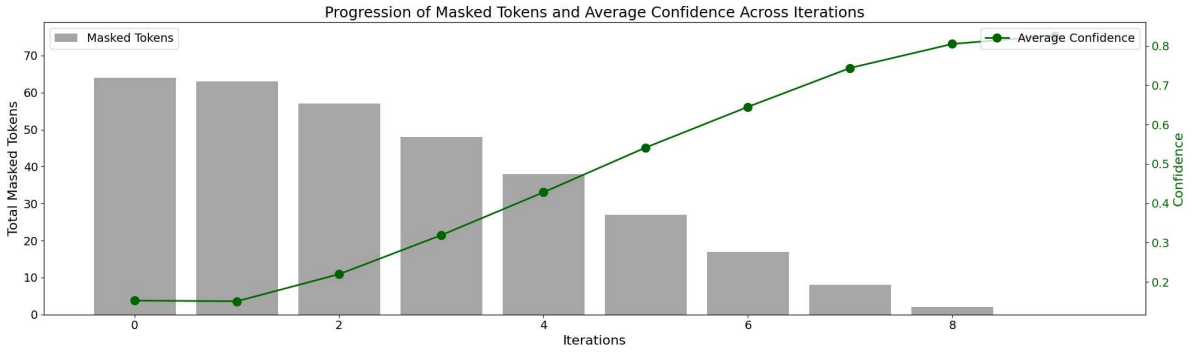

Figure 4. Progression of Masked Tokens and Average Confidence Across Iterations. This figure visualizes the iterative refinement process in Confidence-Guided Sampling. The gray bars represent the number of masked tokens at each iteration, starting from a fully masked sequence and progressively decreasing. The green curve shows the corresponding average confidence in the model's predictions, which increases steadily with iterations. This dynamic showcases the effectiveness of the sampling strategy in resolving ambiguities and refining 3D pose estimates, culminating in high-confidence predictions with minimal masking by the final iteration.

L. Impact of Pose Tokenizer on MaskHand

The results presented in Table 9 demonstrate the critical influence of the Pose Tokenizer's design on the performance of MaskHand. Increasing the codebook size from 1024×256 to 2048×256 yields significant improvements in both PA-MPJPE and MVE metrics across the HO3Dv3 and FreiHAND datasets. This indicates that a moderately larger codebook provides richer and more expressive pose representations, enabling better reconstruction of complex 3D hand poses. However, expanding the codebook further to 4096×256 diminishes accuracy, suggesting that an overly large codebook introduces unnecessary complexity, making it harder for the model to generalize effectively.

Table 9. Impact of Codebook Size (Tokens = 96) on MaskHand.

	HO3Dv3		FreiHAND	
# of code × code dimension	MPJPE (↓)	MVE (↓)	MPJPE (↓)	MVE (↓)
1024×256	7.2	7.3	6.1	6.1
2048×128	7.1	7.2	5.7	5.6
2048×256	7.0	7.0	5.5	5.4

M. Impact of Multi-Scale Features

The ablation study on multi-scale feature resolutions in MaskHand (as shown in Table 10) highlights the trade-off between accuracy and computational cost. Including resolutions up to $4\times$ yields slight accuracy gains, with PA-MPJPE reducing to 7.0 mm on HO3Dv3 and 5.7 mm on FreiHAND. However, the addition of higher resolutions, such as $1\times$ and $2\times$, results in inconsistent or degraded performance. Specifically, the inclusion of $1\times, 4\times, 8\times$ scales increases PA-MPJPE to 7.2 mm on HO3Dv3 and 5.8 mm on FreiHAND. Adding $2\times$ further worsens performance, reaching 7.6 mm on HO3Dv3 and 6.1 mm on FreiHAND, while significantly increasing computational overhead. Notably, the omission of lower-scale features (e.g., $1\times$) leads to performance degradation, highlighting the importance of combining fine-grained details with holistic structure. While multi-scale features remain critical, the study demonstrates that not all resolutions contribute equally, with $1\times$ and $4\times$ emerging as the optimal balance for accuracy and computational efficiency.

Table 10. Impact of feature resolutions on PA-MPJPE and PA-MPVPE errors for HO3Dv3 and FreiHAND datasets.

Feature Scales (Included)	**HO3Dv3 ↓**		**FreiHAND ↓**	
	PA-MPJPE	**PA-MPVPE**	**PA-MPJPE**	**PA-MPVPE**
1×	7.1	7.1	5.6	5.6
1×, 4×	7.0	7.0	5.5	5.4
1×, 4×, 8×	7.2	7.2	5.6	5.7
1×, 8×, 16×	7.1	7.1	5.7	5.6
1×, 4×, 8×, 16×	7.6	7.5	5.9	6.0

N. Deformable Cross-Attention Layers in MaskHand

The ablation study in Table 11 highlights the pivotal role of Deformable Cross-Attention Layers in the Context-Infused Masked Synthesizer of MaskHand. Increasing layers from 2 to 4 yields significant performance gains, reducing PA-MPJPE to 5.7 mm and PA-MPVPE to 5.5 mm on the FreiHAND dataset. This improvement underscores the layers' effectiveness in fusing multi-scale contextual features and refining token dependencies for enhanced 3D hand mesh reconstruction. However, further increasing the layers beyond 4 results in diminishing returns, with slight performance degradation at 6 and 8 layers (e.g., PA-MPVPE increases to 5.8 mm and 6.1 mm, respectively). This decline suggests that additional layers introduce unnecessary complexity, potentially overfitting or disrupting the model's ability to generalize effectively. The findings

reveal that 4 layers provide the optimal balance, leveraging the benefits of cross-attention mechanisms without incurring computational overhead or accuracy trade-offs.

Table 11. Impact of Deformable Cross Attention Layers on FreiHAND dataset.

Metric	2 Layers	4 Layers	6 Layers	8 Layers
PA-MPJPE	6.3	5.5	**5.5**	5.7
PA-MPVPE	6.7	**5.4**	5.8	5.8

O. Qualitative Results in the Wild

Comparison of State-of-the-Art (SOTA) Methods. Figure 5 demonstrates the superiority of MaskHand over other SOTA methods in recovering 3D hand meshes. Unlike competing approaches, MaskHand employs a generative masked modeling framework, enabling it to synthesize unobserved or occluded hand regions. This capability allows MaskHand to achieve robust and precise 3D reconstructions, even in scenarios with heavy occlusions, intricate hand-object interactions, or diverse hand poses. By refining masked tokens, MaskHand effectively addresses ambiguities in the 2D-to-3D mapping process, resulting in highly accurate reconstructions.

Multiple Reconstruction Hypotheses with Explicit Confidence Levels. Figure 6 and 7 illustrates MaskHand's 3D hand mesh reconstructions in occluded scenarios, ranked by confidence. The comparison of reconstructions across different confidence levels reveals that high-confidence hypotheses produce meshes that closely align with the ground truth, ensuring structural accuracy and fidelity. As confidence decreases (e.g., from the 100th to the 1000th hypothesis), the reconstructions degrade, exhibiting distortions and unrealistic poses. This highlights the significance of MaskHand's confidence-aware modeling, where prioritizing high-confidence hypotheses leads to more accurate and robust 3D hand reconstructions.

Reference Key Points in the Deformable Cross-Attention. Figure 8 visualizes the interaction between reference keypoints (yellow) and sampling offsets (red) in the Deformable Cross-Attention module of MaskHand's Masked Synthesizer. By leveraging 2D pose as guidance, the model dynamically samples and refines critical features for accurate 3D reconstruction. This mechanism proves crucial in handling severe occlusions, intricate hand-object interactions, and complex viewpoints, ensuring precise alignment between 2D observations and 3D predictions.

MaskHand's Performance on In-the-Wild Images. Figure 9 highlights MaskHand's robustness in real-world conditions. The model demonstrates its ability to recover accurate 3D hand meshes from single RGB images, excelling in challenging scenarios such as occlusions, hand-object interactions, and diverse hand appearances. This versatility underscores MaskHand's applicability to real-world tasks, where robust and reliable performance is essential. **Challenging Poses from the HInt Benchmark**. Figure 10 and 11 illustrates MaskHand's effectiveness in reconstructing 3D hand meshes for challenging poses from the HInt Benchmark [11]. The model accurately handles extreme articulations and unconventional hand configurations, showcasing its ability to generalize to complex datasets and produce high-fidelity results.

References

[1] Kaggle asl alphabet. Available online: https://www.kaggle.com/grassknoted/asl-alphabet (accessed on 19 July 2021). 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1

[4] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 5, 6

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 645–654, 2024. 5

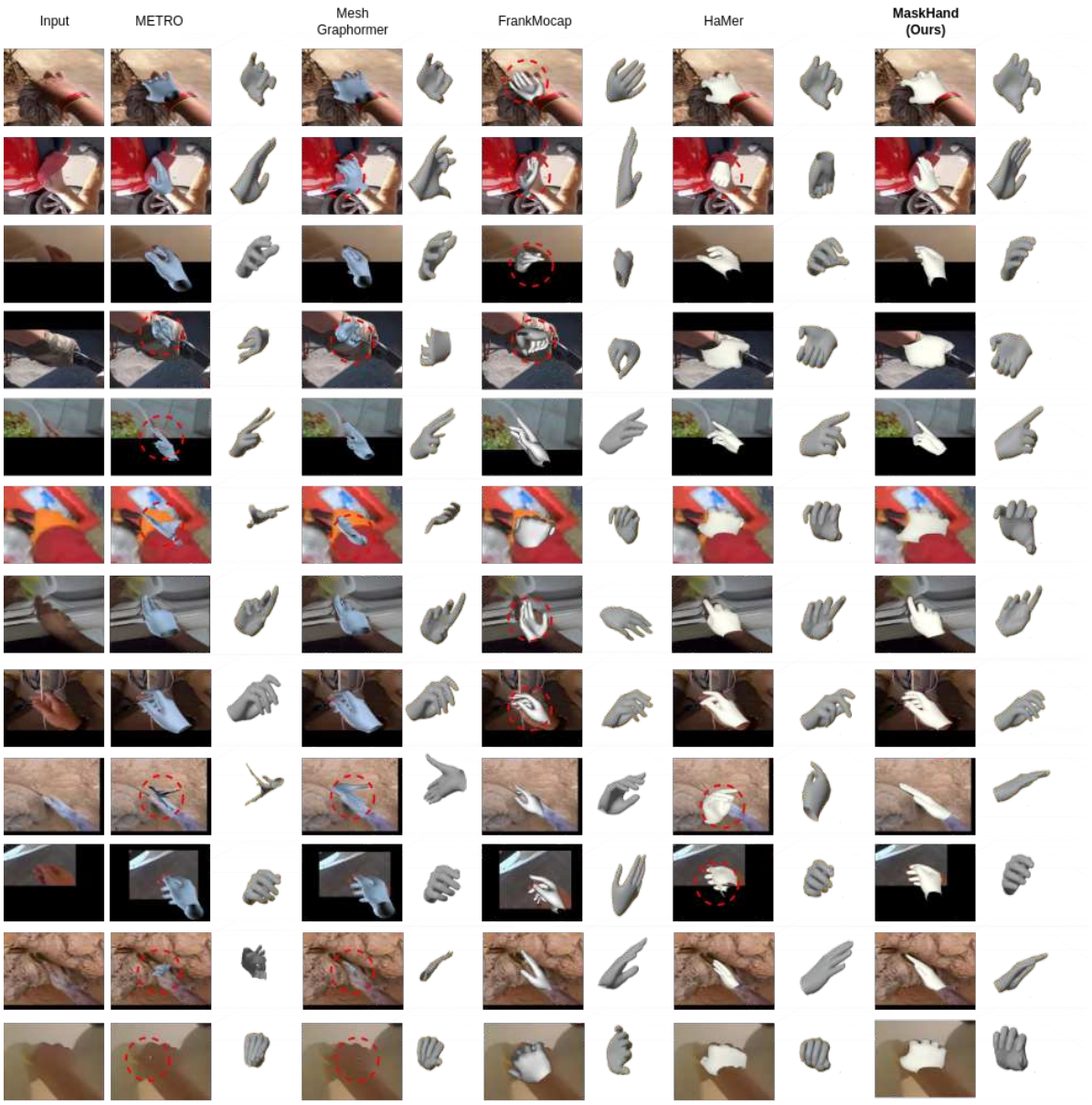

Figure 5. Comparison of State-of-the-Art (SOTA) methods for 3D hand mesh recovery, highlighting the performance of MaskHand. Unlike other approaches, MaskHand employs a generative masked modeling framework to synthesize unobserved or occluded regions, enabling precise and robust 3D hand reconstructions even in challenging scenarios such as heavy occlusions, hand-object interactions, and diverse hand poses. This comparison underscores MaskHand's ability to outperform competing methods by addressing ambiguities in the 2D-to-3D mapping process through its innovative masked token refinement strategy.

[7] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 4

[8] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 4

[9] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 1

[10] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh

estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 4

[11] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 3, 4, 9, 16, 17

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[13] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 4

[14] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 1

[15] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 1

[16] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2, 4, 6

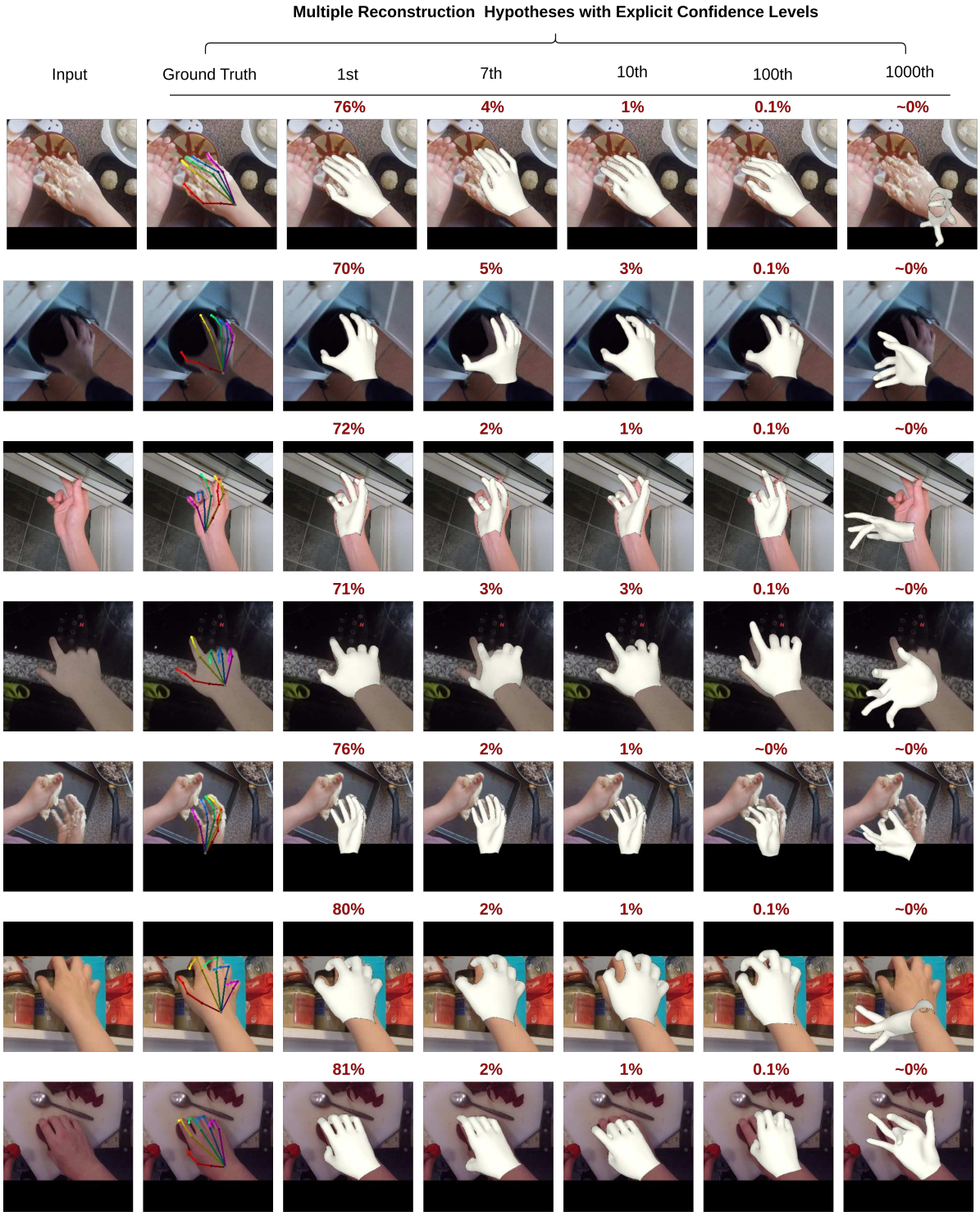

Figure 6. Multiple reconstruction hypotheses with explicit confidence levels. The figure illustrates MaskHand's 3D hand mesh reconstructions in occluded scenarios, ranked by confidence. High-confidence hypotheses closely align with the ground truth, ensuring structural accuracy and fidelity. As confidence decreases (e.g., from the 100th to the 1000th hypothesis), reconstructions degrade, exhibiting distortions and unrealistic poses. This highlights the importance of prioritizing high-confidence hypotheses for robust and accurate 3D hand reconstruction.

Figure 7. Multiple reconstruction hypotheses with explicit confidence levels. The figure illustrates MaskHand's 3D hand mesh reconstructions in occluded scenarios, ranked by confidence. High-confidence hypotheses closely align with the ground truth, ensuring structural accuracy and fidelity. As confidence decreases (e.g., from the 100th to the 1000th hypothesis), reconstructions degrade, exhibiting distortions and unrealistic poses. This highlights the importance of prioritizing high-confidence hypotheses for robust and accurate 3D hand reconstruction.

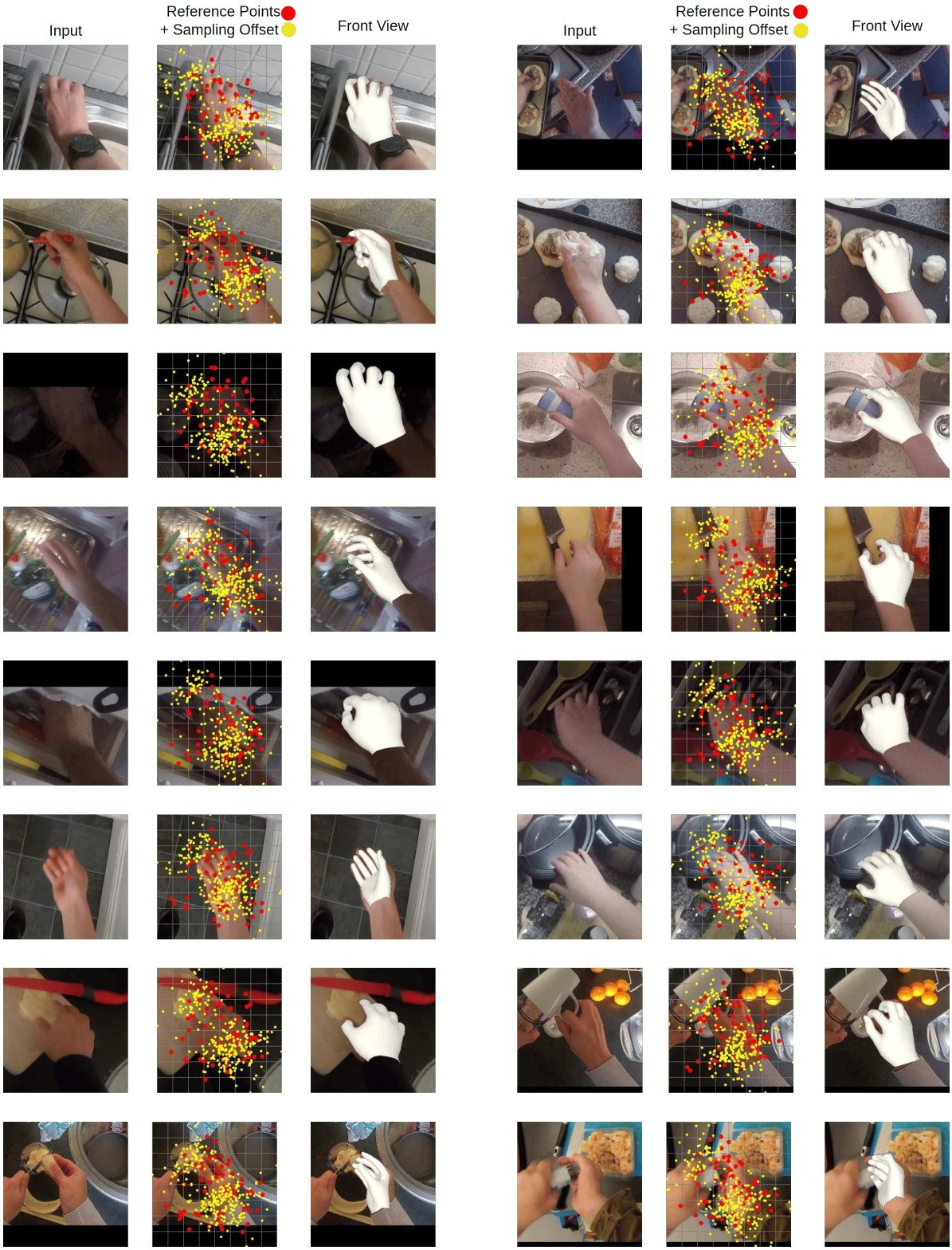

Figure 8. Visualization of reference key points (yellow) and sampling offsets (red) in the Deformable Cross-Attention module of MaskHand's Masked Synthesizer. The 2D pose acts as a guidance signal, enabling the model to dynamically sample and refine features critical for reconstructing accurate 3D hand meshes (rightmost column). This mechanism adapts to challenging scenarios, such as severe occlusions, intricate hand-object interactions, and complex viewpoints, ensuring precise alignment between 2D observations and 3D predictions for robust and high-fidelity reconstructions.

Figure 9. MaskHand's performance on in-the-wild images, demonstrating its ability to recover accurate and robust 3D hand meshes from single RGB inputs. The model excels in challenging scenarios, including occlusions, hand-object interactions, and diverse hand appearances, showcasing its versatility and reliability in real-world conditions.

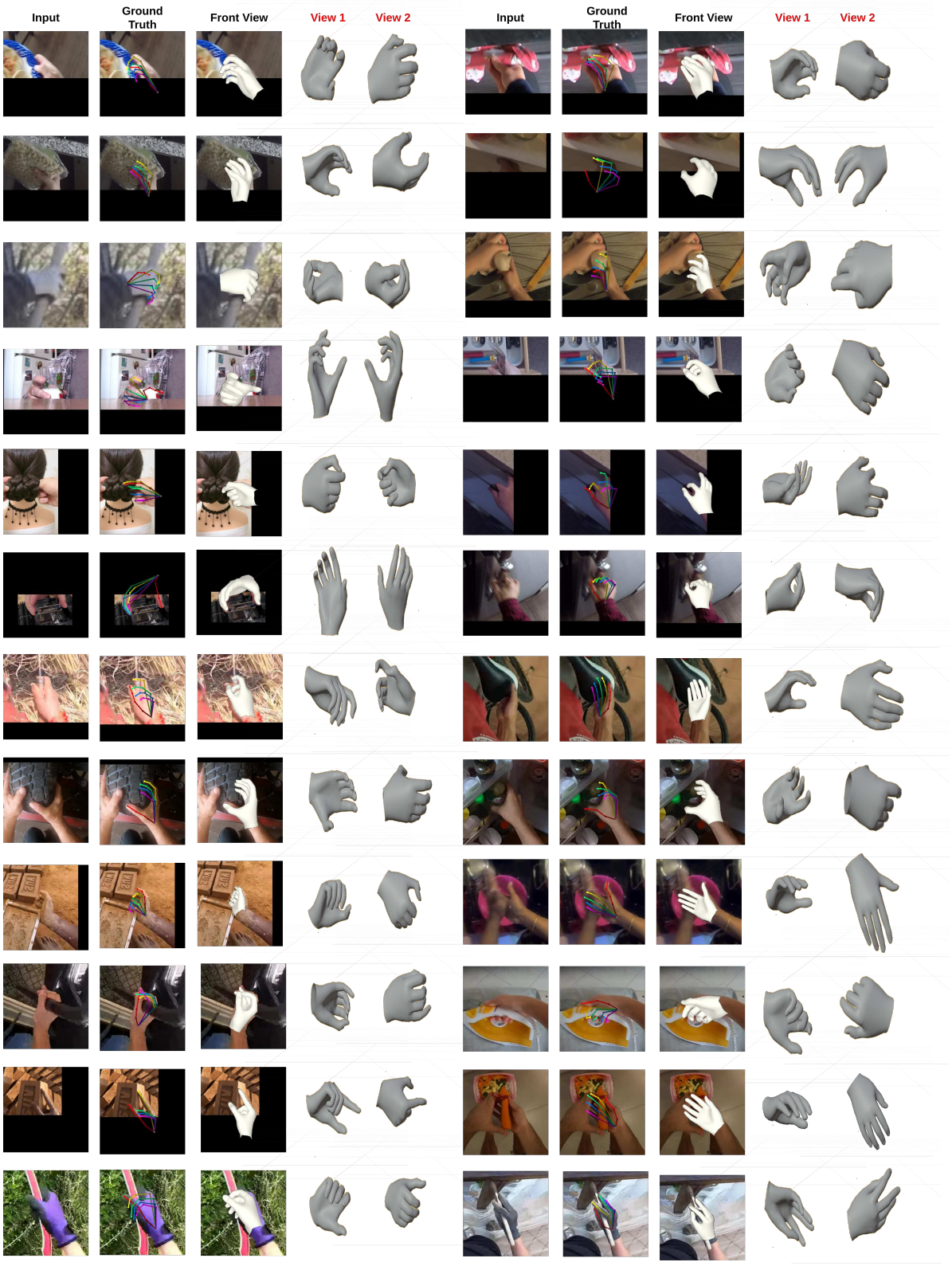

Figure 10. Qualitative results of our approach on challenging poses from the HInt Benchmark [11] dataset.

Figure 11. Qualitative results of our approach on challenging poses from the HInt Benchmark [11] dataset.