CVPR
#5785

MMHMR: Generative Masked Modeling for Hand Mesh Recovery

Anonymous CVPR submission

Paper ID 5785

Figure 1. MMHMR: a novel generative masked model for accurate and robust 3D hand mesh recovery from single RGB images, excelling in diverse scenarios like occlusions, hand-object interactions, and varied appearances. Watch our Supplemental Video to see it in action!

Abstract

*Reconstructing a 3D hand mesh from a single RGB image is challenging due to complex articulations, self-occlusions, and depth ambiguities. Traditional discriminative methods, which learn a deterministic mapping from a 2D image to a single 3D mesh, often struggle with the inherent ambiguities in 2D-to-3D mapping. To address this challenge, we propose **MMHMR**, a novel generative masked model for hand mesh recovery that synthesizes plausible 3D hand meshes by learning and sampling from the probabilistic distribution of the ambiguous 2D-to-3D mapping process. MMHMR consists of two key components: (1) a **VQ-MANO**, which encodes 3D hand articulations as discrete pose tokens in a latent space, and (2) a **Context-Guided Masked Transformer** that randomly masks out pose tokens and learns their joint distribution, conditioned on corrupted token sequence, image context, and 2D pose cues. This learned distribution facilitates confidence-guided sampling during inference, producing mesh reconstructions with low uncertainty and high precision. Extensive evaluations on benchmark and real-world datasets demonstrate that MMHMR achieves state-of-the-art accuracy, robustness, and realism in 3D hand mesh reconstruction. Project website: https://anonymous-ml-model.github.io/MMHMR/.*

1. Introduction

Hand mesh recovery has gained significant interest in computer vision due to its broad applications in fields such as robotics, human-computer interaction [42, 52], animation, and AR/VR [9, 30]. While previous methods have explored markerless, image-based hand understanding, most depend on depth cameras [2, 22, 43, 47, 53] or multi-view images [4, 23, 50, 51]. Consequently, most of these methods are not feasible for real-world applications where only monocular RGB images are accessible. On the other hand, monocular hand mesh recovery from a single RGB image, especially without body context or explicit camera parameters, is highly challenging due to substantial variations in hand appearance in 3D space, frequent self-occlusions, and complex articulations.

Recent advances, especially in transformer-based methods, have shown significant promise in monocular hand mesh recovery (HMR) by capturing intricate hand structures and spatial relationships. For instance, METRO [12] and MeshGraphormer [39] utilize multi-layer attention mechanisms to model both vertex-vertex and vertex-joint interactions, thereby enhancing mesh fidelity. Later, HaMeR [46] illustrated the scaling benefits of large vision transformers and extensive datasets for HMR, achieving improved reconstruction accuracy. However, these methods are inherently discriminative, producing deterministic out-

puts for each image. Consequently, they face limitations in complex, real-world scenes where ambiguities arise due to occlusions, hand-object interactions, and challenging viewpoints are prevalent.

To overcome these limitations, we introduce MMHMR, a novel generative masked model designed for accurate 3D hand mesh recovery. By learning and sampling from the underlying joint distribution of hand articulations and image features, our model synthesizes most probable 3D hand meshes, mitigating ambiguities inherent in single-view reconstruction. MMHMR consists of two main components: a VQ-MANO and a context-guided masked transformer, which are trained in two consecutive stages. In the first stage, VQ-MANO is trained with Vector Quantized Variational Autoencoders (VQ-VAE) [55] to encode continuous hand poses (e.g., joint rotations) of the MONO parametric hand model into a sequence of discrete pose tokens. In the second stage, the token sequence is partially masked and the context-guided masked transformer is trained to reconstruct the masked tokens by learning token conditional distribution, based on multiple contextual clues, including corrupted token sequence, image features, and 2D pose structures.

This generative masked training allows MMHMR to learn an explicit probabilistic mapping from 2D images to plausible 3D hand meshes. Such probabilistic mapping enables confidence-guided sampling during inference, where only pose tokens with high prediction confidence are retained. This sampling process enables MMHMR to leverage measurable uncertainty to mitigate ambiguities in the 2D-to-3D mapping, resulting in enhanced mesh reconstruction accuracy. Our contributions are summarized as follows

- MMHMR is the first to leverage generative masked modeling for reconstructing robust 3D hand mesh. The main idea is to explicitly learn the 2D-to-3D probabilistic mapping and synthesize high-confidence, plausible 3D meshes by sampling from learned probabilistic distribution.
- We design context-guided masked transformer that effectively fuses multiple contextual clues, including 2D pose, image features, and unmasked 3D pose tokens.
- We propose differential masked training to learn hand pose token distribution, conditioned on all contextual clues. This learned distribution facilitates confidence-guided sampling during inference, producing mesh reconstructions with low uncertainty and high precision.
- We demonstrate through extensive experiments that MMHMR outperforms SOTA methods on standard datasets.

2. Related Work

2.1. Discriminative Methods

Human hand recovery has been developed in recent years, with early approaches [21, 28, 54, 57, 67, 69] leveraging optimization techniques to estimate hand poses based on 2D skeleton detections. Later, MANO [48] introduced a differentiable parametric mesh model that capture hand shape and articulation, allowing the model to provide a plausible mesh with end-to-end estimation of model parameters directly from a single-view image. Boukhayma et al. [5] presented the first fully learnable framework to directly predict the MANO hand model parameters [48] from RGB images. Similarly, several subsequent methods have leveraged heatmaps [66] and iterative refinement techniques [3] to ensure 2D alignment. Kulon et al. [35, 36] proposed a different regression approach that predicts 3D vertices instead of MANO pose parameters, achieving notable improvements over prior methods. Recent methods, such as METRO [12], MeshGraphormer [39], HaMeR [46] achieve the SOTA performance by modeling both vertex-vertex and vertex-joint interactions. These existing methods, based on discriminative regression, learn a deterministic mapping from the input image to the output mesh. This deterministic approach struggles to capture the uncertainties and ambiguities caused by hand self-occlusions, interactions with objects, and extreme poses or camera angles, resulting in unrealistic hand mesh reconstructions.

2.2. Generative Methods

Our MMHMR employs a generative method that learns a probabilistic mapping from the input image to the output mesh. It utilizes this learned distribution to synthesize high-confidence, plausible 3D hand meshes based on 2D visual contexts. HHMR [37] is the only other generative hand mesh recovery method in the literature. Unlike HHMR [37] that utilizes diffusion models, MMHMR is inspired by the success of masked image and language models for image and text generation tasks [6, 7, 15, 16, 68]. This fundamental difference allows MMHMR to explicitly and quantitatively estimate confidence levels or prediction probabilities for all mesh reconstruction hypotheses, enabling confidence-guided hypothesis selection for accurate reconstruction. In contrast, HHMR's denoising diffusion process synthesizes multiple mesh hypotheses without associating a confidence level with each hypothesis. Thus, it only reports the theoretically best mesh reconstruction by finding the hypothesis with minimal reconstruction errors under the assumption that ground-truth meshes are available.

3. Proposed Method: MMHMR

Problem Formulation. Given a single-hand image I, we aim to learn a mapping function $f(I) = \{\theta, \beta, \pi\}$ that re-

gresses the MANO [48] model parameters from the input image. This mapping function encompasses three key components: the hand pose parameters $\theta \in \mathbb{R}^{48}$, shape parameters $\beta \in \mathbb{R}^{10}$, and camera parameters $\pi \in \mathbb{R}^{3}$, enabling comprehensive 3D hand reconstruction.

Overview of Proposed Method. As depicted in Figure 2, the MMHMR architecture comprises two main modules: *VQ-MANO* and the *Context-Guided Masked Transformer*. The process initiates with *VQ-MANO*, which converts 3D MANO pose parameters (θ) into discrete pose tokens. These tokens are then processed by the *Context-Guided Masked Transformer*, which includes a image encoder and masked decoder. The encoder extracts multi-scale image features, which, along with 2D pose guidance and unmasked pose tokens, are fused by masked graph transformer decoder. Within the decoder, *Graph-Guided Pose Modeling* ensures anatomical coherence by modeling joint dependencies, while the *Context-Infused Masked Synthesizer* fuses image features and token dependencies to learn the probabilistic reconstruction of pose tokens via differential masked modeling. During inference, the model iteratively refines pose predictions, retaining high-confidence tokens and re-masking those with low confidence, leveraging image semantics, inter-token relationships, and 2D pose guidance to progressively improve accuracy. Finally, the reconstruction is completed as the predicted pose (θ) , shape (β) and camera parameters (π) are feed into the MANO hand model.

3.1. Hand Model and VQ-MANO

Our approach utilizes the MANO hand model [48], which takes pose parameters $\theta \in \mathbb{R}^{48}$ and shape parameters $\beta \in \mathbb{R}^{10}$ as input. The function $M(\theta, \beta)$ outputs a 3D hand mesh $M \in \mathbb{R}^{V\times 3}$ with $V = 778$ vertices and joint locations $X \in \mathbb{R}^{K\times 3}$ with $K = 21$ joints, enabling both surface and pose representation.

The VQ-MANO is a MANO hand tokenizer that learn a discrete latent space for 3D pose parameters $\theta \in \mathbb{R}^{48}$ by quantizing the continuous pose embeddings into a learned codebook C with discrete code entries, as depicted in Figure 2(a). To this end, we employ a Vector Quantized Variational Autoencoder (VQ-VAE) [55] for pretraining the tokenizer. Specifically, we input the MANO pose parameters θ into a convolutional encoder E, which maps them to a latent embedding z. Each embedding z_i is then quantized to its nearest codebook entry $c_k \in C$ based on Euclidean distance, defined as

$$\hat{z}_i = \arg\min_{c_k \in C} \|z_i - c_k\|_2.$$

The total loss function of VQ-MANO is formulated as:

$$\mathcal{L}_{\text{vq-mano}} = \lambda_{\text{re}}\mathcal{L}_{\text{recon}} + \lambda_{\text{E}}\|\,\text{sg}[z] - c\|_2 + \lambda_{\alpha}\|z - \text{sg}[c]\|_2,$$

where $\mathcal{L}_{\text{vq-mano}}$ consists of a MANO reconstruction loss $\mathcal{L}_{\text{recon}}$, a latent embedding loss, and a commitment loss, weighted by hyperparameters λ_{re}, λ_{E}, and λ_{α}, respectively. Here, $\text{sg}[\cdot]$ denotes the stop-gradient operator, which prevents gradients from flowing through its argument during backpropagation. To further enhance reconstruction quality, we incorporate an additional L1 loss:

$$\mathcal{L}_{\text{recon}} = \lambda_{\theta}\mathcal{L}_{\theta} + \lambda_{\text{V}}\mathcal{L}_{\text{V}} + \lambda_{\text{J}}\mathcal{L}_{\text{J}},$$

which aims to minimize the discrepancies between the predicted and ground-truth MANO parameters, including the pose parameters θ, mesh vertices V, and hand joints J. The tokenizer is optimized using a straight-through gradient estimator to facilitate gradient propagation through the non-differentiable quantization step. Additionally, the codebook entries C are updated via exponential moving averages and periodic codebook resets, as described in [19, 59].

3.2. Context-Guided Masked Transformer

The context-guided masked transformer comprises two main components: the multi-scale image encoder and the masked graph transformer decoder.

3.2.1. Multi-scale Image Encoder

Our encoder uses a vision transformer (ViT-H/16) to extract image features [46], processing 16x16 pixel patches into feature tokens. Following ViTDet [1], we adopt a multi-scale feature approach by upsampling the initial feature map to produce feature maps at varying resolutions. This multi-scale representation is critical for handling complex articulations and self-occlusions in hand poses. High-resolution maps provide fine-grained joint details, while low-resolution maps capture global hand structure, balancing precision in joint positioning with overall anatomical coherence. Moreover, we utilize cross-attention with low-resolution feature maps in the x-Attention head to regress stable shape parameters (β) and camera orientation (π), making the process computationally efficient. This approach decouples shape estimation from pose modeling, preserving morphological stability and spatial alignment, and enhancing robustness and anatomical accuracy in 3D hand reconstruction.

3.2.2. Masked Graph Transformer Decoder

The Masked Transformer Decoder is composed of two key components: Graph-Guided Pose Modeling (GGPM) and the Context-Infused Masked Synthesizer Module.

Graph-Guided Pose Modeling (GGPM). Our decoder employs 2 blocks of lightweight graph transformer that processes pose tokens generated by VQ-MANO, enriched with 2D pose guidance, where hand pose tokens are represented as graph nodes linked by learnable adjacency matrices to capture joint relationships effectively. To enhance stability and anatomical accuracy, we integrate a transformer

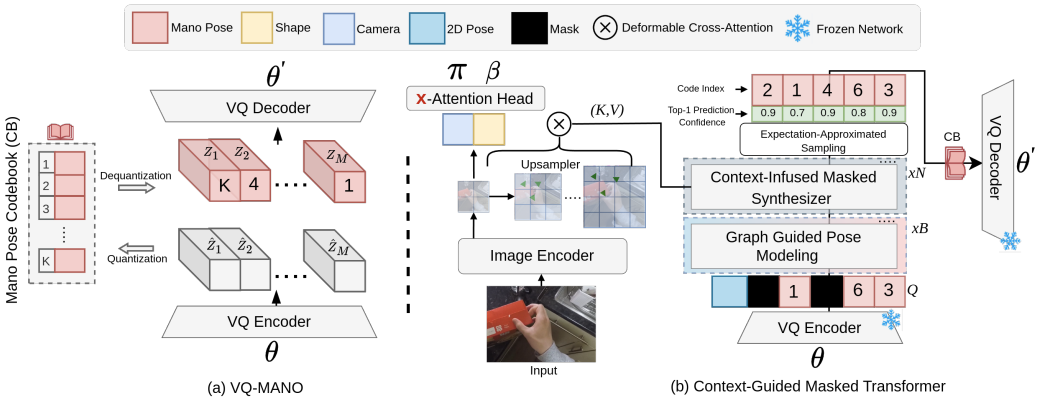

Figure 2. MMHMR *Training Phase*. MMHMR consists of two key components: (1) **VQ-MANO**, which encodes 3D hand poses into a sequence of discrete tokens within a latent space, and (2) a **Context-Guided Masked Transformer** that models the probabilistic distributions of these tokens, conditioned on the input image, 2D pose cues, and a partially masked token sequence.

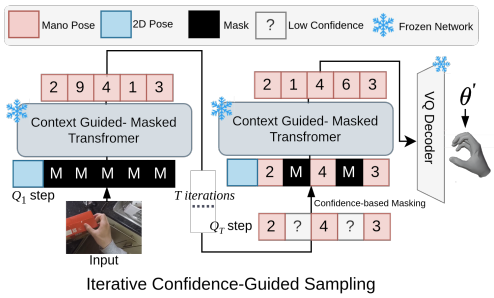

Figure 3. MMHMR *Inference Phase*: **Confidence-Guided Iterative Sampling** — a step-by-step refinement of pose selection by probabilistically sampling high-confidence tokens.

encoder with a Squeeze-and-Excitation (SE) block [29], which emphasizes joint orientations and angles and ensures spatial alignment through a 1x1 convolution layer. Within the transformer encoder, Multi-Head Attention (MHA) and pointwise convolution layers refine high-resolution dependencies between joints, producing a cohesive and anatomically aligned 3D pose representation, denoted as Q_{GGPM}. The output Q_{GGPM} is formulated as:

$$Q' = \text{MHA}(\text{Norm}(Q_{\text{C}})) + \text{Conv}(\text{Norm}(Q_{\text{C}})) + Q_{\text{C}} \quad (1)$$

$$Q_{\text{GGPM}} = \text{SE}(\text{Norm}(Q')) \quad (2)$$

where $Q_{\text{C}} \in \mathbb{R}^{K \times D}$ represents refined MANO pose tokens and 2D pose guidance queries, and Q_{GGPM} is the stabilized, anatomically consistent pose representation. The refined pose token queries Q_{GGPM} are then passed to the Context-Infused Masked Synthesize to synthesize a precise 3D hand mesh that integrates global and local anatomical features.

Context-Infused Masked Synthesizer. We leverage a multi-layer transformer whose inputs are refined pose tokens Q_{GGPM} and cross-attends them with multi-scale feature maps generated by the image encoder. To enhance computational efficiency with high-resolution feature maps, a deformable cross-attention mechanism is employed [71]. This allows each pose token to focus on a selected set of sampling points around a learnable reference point, rather than the entire feature map. By concentrating attention on relevant areas, the model achieves a balance between computational efficiency and spatial precision, preserving essential information for accurate 3D hand modeling. The deformable cross-attention is defined as:

$$\text{MCDA}(Q_{\text{GGPM}}, \hat{p}_y, \{x^l\}) = \sum_{l,k} A_{lyk} \cdot \mathbf{W} x^l(\hat{p}_y + \Delta p_{lyk}) \quad (3)$$

where Q_{GGPM} are refined manopose token queries, $\hat{p}_y$ are learnable reference points, Δp_{lyk} are sampling offsets, $\{x^l\}$ are multi-scale features, A_{lyk} are attention weights, and $\mathbf{W}$ is a learnable weight matrix. With the inclusion of a `[MASK]` token in the masked transformer decoder, the module can predict masked pose tokens during training, while also facilitating token generation during inference. This approach allows the `[MASK]` token to serve as a placeholder for final pose token predictions, supporting robust synthesis of occluded or unobserved hand parts for a coherent 3D hand reconstruction.

3.3. Training: Differential Masked Modeling

Context-conditioned Masked Modeling. We employ masked modeling to train our model that learns the probabilistic distribution of 3D hand poses, conditioned on multiple contextual cues. Given a sequence of discrete pose tokens $Y = [y_i]_{i=1}^{L}$ from the pose tokenizer, where L is the sequence length, we randomly mask out a subset of m tokens with $m = \lceil \gamma(\tau) \cdot L \rceil$, where $\gamma(\tau)$ is a cosine-based masking ratio function. Here, τ is drawn from a uniform distribution $U(0, 1)$, and we adopt the masking func-

tion $\gamma(\tau) = \cos\left(\frac{\pi\tau}{2}\right)$, inspired by generative text-to-image modeling strategies [6].

Masked tokens are replaced with learnable `[MASK]` tokens, forming a corrupted sequence $Y_{\overline{\mathbf{M}}}$ that the model must reconstruct. Each token y_i is predicted based on the probabilistic distribution $p(y_i|Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})$, conditioned on corrupted token sequence $Y_{\overline{\mathbf{M}}}$, 2D pose embedding X_{2D}, and image prompt X_{img}. This approach enables the model to explicitly account for the uncertainty inherent in mapping 2D observations to a coherent 3D hand mesh. The training objective is to minimize the negative log-likelihood of correctly predicting each pose token in the sequence, formulated as follows:

$$\mathcal{L}_{\text{mask}} = -\mathbb{E}_{Y\in\mathcal{D}}\left[\sum_{\forall i\in[1,L]} \log p(y_i|Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})\right].$$

Expectation-Approximated Differential Sampling. The training objective, $\mathcal{L}_{\text{mask}}$, captures stochastic uncertainty in hand mesh reconstruction, enabling precise estimation of the pose parameter θ within a structured discrete latent space. Recent studies [46] demonstrate that applying auxiliary 3D and 2D joint losses—measuring alignment between predicted and ground-truth joints in both 3D coordinates and their 2D projections—further refines pose recovery. In generative masked model training, integrating these auxiliary losses requires transforming latent pose tokens into the MANO pose parameter θ, a process that involves non-differentiable probabilistic sampling. To address this, instead of sampling the distribution $p(y_i|Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})$ to obtain the most probable token from codebook, we implement an expectation-based differential relaxation: instead of directly estimating discrete code indices, the model outputs logits L for each token. These logits undergo a softmax operation, producing the token distribution, which are multiplied by the pretrained codebook, resulting in the mean quantized feature representations $\bar{z} = [\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_M]$:

$$\bar{z} = \text{softmax}(L_{M\times K}) \times \text{CB}_{K\times D}$$

where L represents the logits matrix, CB denotes the codebook, M is the token count, K specifies the codebook size, and D defines the dimensionality of each codebook entry. This mean token embeddings $\bar{z}$ are feed into the decoder to reconstruct the pose parameter θ'. Combined with shape parameters β and camera parameters π, this process reconstructs the 3D hand mesh, enabling the computation of both 3D joint loss and 2D projection loss. Since obtaining $\bar{z}$ is differential, the model can be trained end-to-end using the overall loss function

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

which combines the masked token prediction loss ($\mathcal{L}_{\text{mask}}$), 3D joint loss ($\mathcal{L}_{\text{3D}}$), 2D projection loss ($\mathcal{L}_{\text{2D}}$), and MANO parameter loss $\mathcal{L}_{\text{MANO}}$. Together, these terms collectively minimize discrepancies in the shape and pose parameters within the MANO space, leading to a precise reconstruction of the 3D hand mesh.

3.4. Inference: Confidence-Guided Sampling

Our model leverages confidence-guided sampling to achieve precise and stable 3D hand pose predictions. This process begins with a fully masked sequence Y_1 of length L, with each token initialized as `[MASK]`. Over T decoding iterations, each iteration t applies stochastic sampling to predict masked tokens based on their distributions $p(y_i|Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})$. Following each sampling step, tokens with the lowest prediction confidences are re-masked to be re-predicted in subsequent iterations. The number of tokens re-masked is determined by a masking schedule $\lceil\gamma\left(\frac{t}{T}\right)\cdot L\rceil$, where γ is a decaying function of t. This schedule dynamically adjusts masking intensity based on confidence, using a higher masking ratio in earlier iterations when prediction confidence is lower. Consequently, the model iteratively refines ambiguous regions, progressively improving prediction confidence as context builds. The masking ratio decreases with each step, stabilized by the cosine decay function γ; alternative decay functions are discussed in the supplementary material.

4. Experiments

Datasets.To train the hand pose tokenizer, we employed a diverse set of datasets to capture a wide range of hand poses and interactions. This includes DexYCB [8], InterHand2.6M [44], MTC [60], and RHD [72]. For training MMHMR, we utilized a diverse dataset, following a similar setup as in [46] to ensure a fair comparison. Specifically, the training data was drawn from FreiHAND [73], HODv2 [25], MTC [60], RHD [72], InterHand2.6M [44], H2O3D [25], DexYCB [8], COCO-Wholebody [32], Halpe [20], and MPII NZSL [50].

Evaluation Metrics. Following standard protocols [37, 46, 70], MMHMR evaluated on reconstructed 3D joints using PA-MPJPE and AUC_J, while 3D mesh vertices were evaluated with PA-MPVPE, AUC_V, F@5mm, and F@15mm. Additionally, to examine MMHMR's generalization and accuracy in diverse real-world settings, we employed the Percentage of Correct Keypoints (PCK) [46] metric at multiple thresholds, ensuring a robust evaluation of its performance across varied conditions. To evaluate hand image generation with mesh-guided control, we compute FID-H and KID-H on cropped hand regions and use MediaPipe [64] as a hand detector to measure confidence.

3D Reconstruction Accuracy Evaluation.To comprehensively evaluate MMHMR's 3D joints and mesh reconstruc-

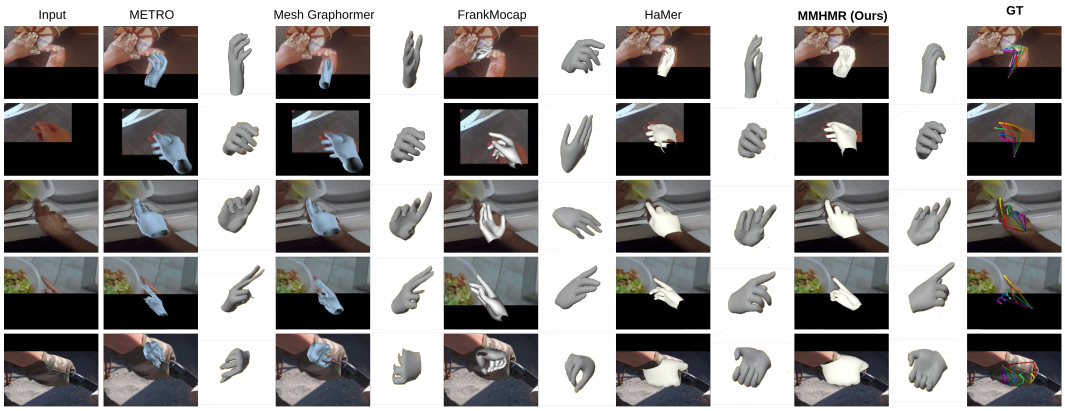

Figure 4. Comparison of State-of-the-Art (SOTA) Methods: Our method, MMHMR, synthesizes unobserved parts through generative modeling, enabling accurate 3D reconstructions in complex, occluded scenarios.

Table 1. 3D Reconstruction Accuracy Evaluation on the HO3Dv3 [26] Benchmark: Comparison with SOTA Methods. "-" indicates metrics not reported by the model.

Method	**Venue**	PA-MPJPE (↓)	PA-MPVPE (↓)	F@5mm (↑)	F@15mm (↑)	AUC_J (↑)	AUC_V (↑)
S^2HAND [10]	*CVPR 2021*	11.5	11.1	0.448	0.932	0.769	0.778
KPT-Transf. [26]	*CVPR 2022*	10.9	-	-	-	0.785	-
ArtiBoost [61]	*CVPR 2022*	10.8	10.4	0.507	0.946	0.785	0.792
Yu et al. [63]	*BMVC 2022*	10.8	10.4	-	-	-	-
HandGCAT [58]	*ICME 2022*	9.3	9.1	0.552	0.956	0.814	0.818
AMVUR [31]	*CVPR 2023*	8.7	8.3	0.593	0.964	0.826	0.834
HMP [17]	*WACV 2024*	10.1	-	-	-	-	-
SPMHand [41]	*TMM 2024*	8.8	8.6	0.574	0.962	-	-
MMHMR	***Ours***	**7.0**	**7.0**	**0.663**	**0.984**	**0.860**	**0.860**

Table 2. 3D Reconstruction Accuracy on the FreiHAND [73] Benchmark: Comparison with State-of-the-Art Methods. †Indicates use of Test-Time Augmentation (TTA). PA-MPJPE and PA-MPVPE are reported in millimeters (mm). "-" indicates metrics not reported by the model.

Methods	**Venue**	PA-MPJPE (↓)	PA-MPVPE (↓)	F@5mm (↑)	F@15mm (↑)
MeshGraphormer† [39]	*ICCV 2021*	5.9	6.0	0.764	0.986
FastMETRO [13]	*ECCV 2022*	6.5	7.1	0.687	0.983
FastViT [56]	*ICCV 2023*	6.6	6.7	0.722	0.981
AMVUR [31]	*CVPR 2023*	6.2	6.1	0.767	0.987
Deformer [62]	*CVPR 2023*	6.2	6.4	0.743	0.984
PointHMR [34]	*CVPR 2023*	6.1	6.6	0.720	0.984
Zhou et al. [70]	*CVPR 2024*	5.7	6.0	0.772	0.986
HaMeR [46]	*CVPR 2024*	6.0	5.7	0.785	0.990
HHMR§ [37]	*CVPR 2024*	5.8	5.8	-	-
MMHMR‡	***Ours***	**5.7**	**5.5**	**0.793**	**0.991**

‡ **MMHMR** reports most confident mesh reconstruction, guided by the learned 2D-to-3D distribution **without** assumption of available GT meshes.
§ HHMR reports best hypothesis mesh reconstruction with minimum MPJPE and MPVPE, **with** the assumption of available GT meshes.

tion capabilities, we used the HO3Dv3 [26] and FreiHAND datasets. HO3Dv3 [26], the largest 3D hand pose test dataset with 20K annotated images, serves as a rigorous test ground with diverse hand poses and complex scenarios. MMHMR was directly tested on HO3Dv3 without prior training on this dataset, assessing its ability to generalize to new data and confirming the model's robustness in challenging settings. This evaluation underscores MMHMR's resilience to dataset-specific biases, proving its adaptability across varied scenarios. Additionally, we evaluated MMHMR on the FreiHAND dataset [73], which includes 4K images spanning controlled environments. For fair comparison, as previous methods like MeshGraphormer [39] and HHMR [37] were evaluated on FreiHAND, we evaluated MMHMR trained solely on the FreiHAND [73].

Real-World Robustness Evaluation. To validate MMHMR's adaptability, we tested it on the HInt benchmark [46] without prior training. HInt introduces diverse

Table 3. Real-World Robustness Evaluation on the HInt Benchmark [46] 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 [49]	*ICCVW 2021*	16.1	41.4	60.2	16.8	45.6	66.2	13.1	36.9	55.8
METRO [38]	*CVPR 2021*	14.7	38.8	57.3	16.8	45.4	65.7	13.2	35.7	54.3
MeshGraphormer [39]	*ICCV 2021*	16.8	42.0	59.7	19.1	48.5	67.4	14.6	38.2	56.0
HandOccNet (param) [45]	*CVPR 2022*	9.1	28.4	47.8	8.1	27.7	49.3	7.7	26.5	47.7
HandOccNet (no param) [45]	*CVPR 2022*	13.7	39.1	59.3	12.4	38.7	61.8	10.9	35.1	58.9
HaMeR [46]	*CVPR 2024*	48.0	78.0	88.8	43.0	76.9	89.3	38.9	71.3	84.4
MMHMR	**Ours**	**48.7**	**79.2**	**90.0**	**46.1**	**81.4**	**92.1**	**46.4**	**77.5**	**90.1**
Visible Joints										
FrankMocap [49]	*ICCVW 2021*	20.1	49.2	67.6	20.4	52.3	71.6	16.3	43.2	62.0
METRO [38]	*CVPR 2021*	19.2	47.6	66.0	19.7	51.9	72.0	15.8	41.7	60.3
MeshGraphormer [39]	*ICCV 2021*	22.3	51.6	68.8	23.6	56.4	74.7	18.4	45.6	63.2
HandOccNet (param) [45]	*CVPR 2022*	10.2	31.4	51.2	8.5	27.9	49.8	7.3	26.1	48.0
HandOccNet (no param) [45]	*CVPR 2022*	15.7	43.4	64.0	13.1	39.9	63.2	11.2	36.2	56.0
HaMeR [46]	*CVPR 2024*	60.8	87.9	94.4	56.6	88.0	94.7	52.0	83.2	91.3
MMHMR	**Ours**	**61.0**	**87.1**	**94.8**	**62.1**	**90.2**	**95.0**	**59.3**	**88.3**	**94.4**
Occluded Joints										
FrankMocap [49]	*ICCVW 2021*	9.2	28.0	46.9	11.0	33.0	55.0	8.4	26.9	45.1
METRO [38]	*CVPR 2021*	7.0	23.6	42.4	10.2	32.4	53.9	8.0	26.2	44.7
MeshGraphormer [39]	*ICCV 2021*	7.9	25.7	44.3	10.9	33.3	54.1	9.3	32.6	51.7
HandOccNet (param) [45]	*CVPR 2022*	7.2	23.5	42.4	7.4	26.1	46.7	7.2	26.1	45.7
HandOccNet (no param) [45]	*CVPR 2022*	9.8	31.2	50.8	9.9	33.7	55.4	9.6	31.1	52.7
HaMeR [46]	*CVPR 2024*	27.2	60.8	78.9	25.9	60.8	80.7	23.0	56.9	76.3
MMHMR	***Ours***	**29.4**	**64.1**	**80.3**	**31.4**	**64.2**	**83.6**	**29.4**	**65.3**	**80.1**

real-world conditions, including varied lighting, angles, and hand-object interactions, offering a realistic evaluation framework. Using the Percentage of Correct Keypoints (PCK) metric, we assessed MMHMR on HInt-NewDays [11], HInt-EpicKitchensVISOR [14], and HInt-Ego4D [24], highlighting its robust generalization and effectiveness in complex environments.

4.1. Comparison to State-of-the-art Approaches

We evaluate MMHMR against a range of state-of-the-art methods (SOTA) on the HO3Dv3 [26], FreiHAND [73], and HInt [46] benchmarks, as detailed in Table 1, Table 2, and Table 3. MMHMR consistently outperforms competing methods across key evaluation metrics, demonstrating robust accuracy in 3D hand reconstruction. Notably, we perform zero-shot evaluations on both the HO3Dv3 and HInt benchmarks to assess MMHMR's generalizability. A core contributor to MMHMR's success is its capability to model and refine uncertainty, making it highly effective in scenarios with complex hand poses and significant occlusions. On the HO3Dv3 dataset (Table 1), MMHMR achieves a PA-MPJPE reduction of approximately 19.5% and a PA-MPVPE reduction of 15.7% compared to the existing SOTA method. This improvement underscores MMHMR's precision in handling challenging hand poses and occlusions. Similarly, on the HInt benchmark (Table 3), MMHMR achieves notable improvements in occluded joint reconstruction at the PCK@0.05 threshold on the HInt benchmark. Specifically, MMHMR shows an 8.1% increase on HInt-NewDays, a 21.2% increase on HInt-VISOR, and a 27.8% increase on HInt-Ego4D compared to the closest SOTA methods. These gains underscore MMHMR's effectiveness in synthesizing unobserved parts and handling real-world occlusions with high accuracy. This highlights MMHMR's strength in synthesizing unobserved parts, enabling accurate reconstruction of occluded regions.

Mesh-Guided Control for Hand Image Generation. As shown in the Generation Pipeline (Figure 5) and Table 5, MMHMR demonstrates adaptability in generating realistic hand images under diverse conditions. We train ControlNet [65]on DexYCB [8] and test it on HO3Dv3 [26]. MMHMR processes input images through a mesh estimator i.e. (MMHMR) to extract 3D hand meshes, which serve as control signals in ControlNet [65] alongside text prompts. Compared to HaMer [46] as a mesh estimator method [46], MMHMR achieves lower FID-H and KID-H scores, indicating better image quality, and a higher hand detection confidence score, reflecting improved anatomical accuracy. Accurate mesh estimation ensures precise control signals for consistent, high-quality hand generation. More details are in the supplementary material.

Table 4. Mesh-Guided Control for Hand Image Generation

Mesh Estimator	FID-H ↓	KID-H ↓	Hand Det Conf. ↑
HaMer [46]	45.65	0.035	0.865
MMHMR (Ours)	40.23	0.027	0.912

4.2. Ablation Study

The effectiveness of MMHMR is grounded in its mask modeling and iterative decoding techniques. This abla-

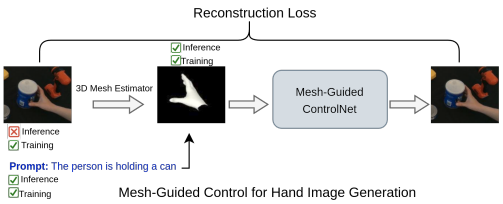

Figure 5. Generation Pipeline: MMHMR extracts 3D hand meshes from images during training, using them with text prompts to guide ControlNet. At inference, only text and mesh signals are needed to generate realistic hand images.

tion study examines how iterative refinement and mask-scheduling strategies impact model performance. Additional experimental results and visualizations in the **Supplementary Material** provide in-depth analyses of the key factors driving MMHMR's performance. These include (1) architectural components, such as the VQ-MANO tokenizer, GGPM design, Context-Infused Masked Synthesizer, and feature resolution choices, (2) training strategies, including mask scheduling and regularization via keypoint and MANO losses, and (3) the model's limitations.

Effectiveness of Proposed Components. Our ablation study on the HO3Dv3 dataset [26] (Table 5) underscores the importance of MMHMR's components. The Upsampler plays a key role in enhancing fine-grained details and resolution, while the 2D OpenPose Context provides essential spatial cues; their absence leads to reduced accuracy. The x-Attention Head is crucial for maintaining hand shape stability by integrating shape and camera parameters, with its removal resulting in a decline in 3D accuracy. Iterative Decoding is vital for refining predictions by leveraging contextual cues, with its absence causing significant performance degradation. The largest performance drop is observed when Graph-Guided Pose Modeling (GGPM) is removed, emphasizing its critical role in capturing joint dependencies and ensuring anatomical coherence.

Table 5. Ablation study of testing results on the HO3Dv3 dataset [26] to evaluate the impact of proposed components. 'w/o' denotes 'without'.

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. x-Attention Head	7.3	7.1	0.655	0.855	0.858
w/o. GGPM	7.3	7.3	0.645	0.853	0.854
w/o. Iterative Decoding	7.2	7.2	0.654	0.857	0.857
MMHMR (Full)	**7.0**	**7.0**	**0.663**	**0.860**	**0.860**

Impact of Iterative Confidence-Guided Sampling. The number of iterations in confidence-guided sampling impacts both accuracy and efficiency in 3D hand pose estimation. As shown in Table 6, increasing iterations from 1 to 5 improves accuracy, with PA-MPVPE dropping from 7.2 to 7.0 on HO3Dv3 and from 5.8 to 5.5 on FreiHAND. This refinement enhances precision by leveraging contextual cues and 2D pose guidance. However, increasing to 10 iterations shows minimal benefit, making 5 iterations optimal for balancing accuracy and computational cost.

Table 6. Iterations in Iterative Confidence-Guided Sampling

	HO3Dv3		FreiHAND	
# of iter.	PA-MPJPE	PA-MPVPE	PA-MPJPE	PA-MPVPE
1	7.2	7.2	5.9	5.8
3	7.1	7.1	5.8	5.7
5	7.0	7.0	5.7	5.5
10	7.1	7.0	5.7	5.5

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 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 [26] and FreiHAND [73] 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	6.0	6.1
$\gamma(\tau \in \mathcal{U}(0, 0.5))$	7.1	7.1	5.8	5.7
$\gamma(\tau \in \mathcal{U}(0, 0.7))$	7.0	7.0	5.7	5.5
$\gamma(\tau \in \mathcal{U}(0, 1.0))$	7.2	7.3	6.0	6.0

5. Conclusion

In this paper, we introduced MMHMR, a novel generative masked model designed for accurate and robust 3D hand mesh reconstruction from single RGB images. MMHMR addresses the longstanding challenges posed by complex hand articulations, self-occlusions, and depth ambiguities by leveraging a generative framework that captures and refines hand pose distributions. Central to our approach are two key components: VQ-MANO, which encodes 3D hand articulations as discrete pose tokens in a learned latent space, and the Context-Guided Masked Transformer, which models token dependencies conditioned on both image features and 2D pose cues. This framework employs confidence-guided iterative sampling to refine reconstructions, producing anatomically realistic hand meshes under challenging conditions. MMHMR outperforms state-of-the-art methods, setting a new benchmark for 3D hand mesh modeling with applications in human-computer interaction, AR, and VR.

6. Supplementary Material

A. Overview

The supplementary material is organized into the following sections:

- Section B: Implementation Details

Section B.3: Ablation for VQ-MANO Pose Tokenizer

Section B.4: Effectiveness of Expectation-Approximated Differential Sampling

Section B.3: Confidence-Guided Masking

Section B.3: Pose Token Sampling Techniques

Section B.3: Impact of Control Signals for Hand Image Generation

Section B.3: Impact of VQ-MANO Tokenizer on MMHMR

Section B.3: Ablation of Feature Resolutions

Section B.3: Impact of Deformable Cross-Attention Layers

Qualitative Results

B. Implementation Details

The implementation of MMHMR, 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 [27] 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 [8], InterHand2.6M [44], MTC [60], and RHD [72]. 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 $4\times$, $8\times$, and $16\times$ scales are used to capture both global and local contextual details, allowing the model to handle complex hand articulations and occlusions. The default number of iterations in Confidence-Guided Sampling is 5 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), $\lambda_{\text{3D}} = 5 \times 10^{-2}$, and $\lambda_{\text{2D}} = 1 \times 10^{-2}$. This phase is trained using the Adam optimizer on NVIDIA RTX A6000 GPUs with a batch size of 48 and a learning rate of 1×10^{-5}.

Mesh-Guided Control for Hand Image Generation. To extend the utility of the reconstructed 3D hand meshes, we integrate them as control signals into the publicly available ControlNet framework [65]. This integration enables the generation of high-quality hand images guided by the reconstructed meshes. The ControlNet models are trained on hand-cropped images resized to 256×256, with a batch size of 64, using the Adam optimizer with a learning rate of 1×10^{-5}. Training is conducted over 200 epochs on eight NVIDIA RTX A6000 GPUs. During training, image captions are generated using LLAVA-NEXT [40], a vision-language model capable of producing descriptive captions based on image content. Specifically, we query LLAVA-NEXT with prompts such as "What's in the image and how does the hand pose look like?" to obtain concise descriptions focusing on the hand's pose, including finger positions and hand orientation. For instance, in response to an image of a hand, LLAVA-NEXT might generate a caption like "A right hand with fingers slightly bent, palm facing upward." These captions are then utilized as text guidance during the training of ControlNet. By combining these captions with the corresponding 3D hand meshes extracted using MMHMR, we enable the ControlNet framework to effectively learn a mapping between text, mesh signals, and high-quality hand image outputs. The generation pipeline involves two stages: training and inference. During training, MMHMR extracts 3D hand meshes from images, which are paired with text prompts to guide the ControlNet model. This process ensures that the generated images are both semantically aligned with the textual descriptions and anatomically consistent with the 3D hand meshes. At

CVPR #5785

inference, the model requires only text prompts and mesh signals as inputs. The text prompts provide high-level semantic guidance, such as the desired pose or hand orientation, while the mesh signals ensure that the generated images maintain accurate anatomical and spatial features. This design enables the generation of realistic and contextually appropriate hand images, demonstrating the robustness and flexibility of the proposed approach. By leveraging both text-based and mesh-based guidance, the system achieves high fidelity in image generation, even in scenarios involving complex hand poses or occlusions.

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 MMHMR 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 MMHMR 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 MMHMR 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.

B.3. Ablation for VQ-MANO

Tables 1 and 2 summarize an ablation study on the Freihand [73] 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 pose tokenization.

Table 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	1.01	0.59	0.47	0.41
PA-MPVPE	0.97	0.57	0.44	0.41

Table 2. 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	0.66	0.56	0.47	0.43
PA-MPVPE	0.65	0.58	0.44	0.44

B.4. Effectiveness of Expectation-Approximated Differential Sampling

The results in Table **??** highlight the critical role of Training-Time Differentiable Sampling in enhancing MMHMR's performance, particularly when combined with key regularization losses. Incorporating all losses—L_{mask}, L_θ, L_3D, L_2D, and β—yields optimal performance across all evaluation metrics on both the HO3Dv3 and FreiHAND datasets. The 3D loss (L_3D) maintains structural integrity by aligning predicted 3D joints with ground truth, while the 2D loss (L_2D) addresses monocular reconstruction ambiguities by ensuring alignment between 3D projections and observed 2D keypoints. Notably, excluding either L_3D or L_2D leads to significant increases in errors, underscoring their vital role in producing accurate and plausible 3D reconstructions. These findings demonstrate that differen-

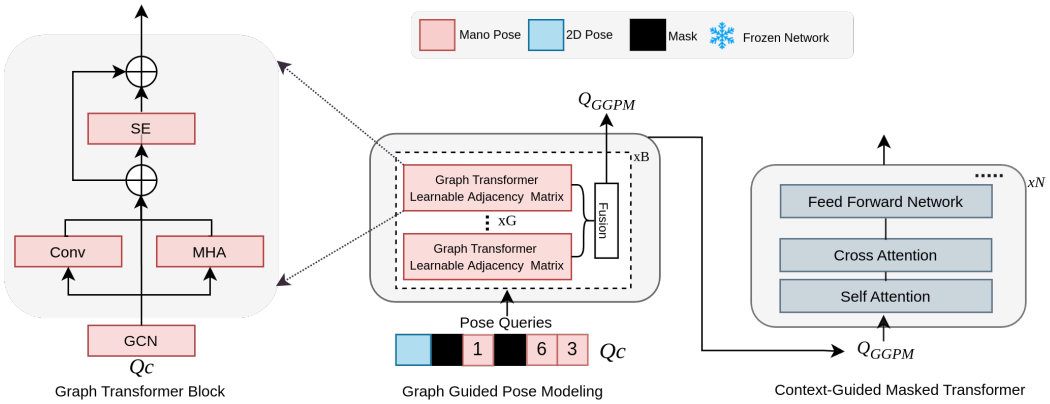

Figure 1. GGPM

tiable sampling not only enables seamless integration of these losses but also ensures their direct contribution to the overall accuracy and reliability of the model's predictions. Our analysis reveals that the combination of differentiable sampling and carefully chosen losses is crucial for achieving high-quality hand mesh recovery. This approach allows the model to effectively learn from various constraints, resulting in more accurate and robust 3D reconstructions across different datasets and evaluation metrics.

B.5. Confidence-Guided Sampling Strategies at Inference

During inference, we employ a Confidence-Guided Sampling approach to iteratively refine pose predictions and address uncertainties in 2D-to-3D pose estimation. The process begins with a fully masked sequence of 64 pose tokens, as illustrated in Figures **??** and 2. Here, the x-axis represents pose token indices (0–64), and the y-axis denotes decoding iterations (0–5). Initially, most tokens are masked (dark green squares), reflecting high uncertainty in early predictions. As the model progresses, the number of masked tokens decreases, replaced by unmasked tokens (light cyan squares) as prediction confidence improves, as shown in Figure **??**. Figure 2 further illustrates the evolution of prediction confidence, transitioning from low (blue) in the initial stages to high (yellow) as the model refines its estimates. Confidence improvements typically start with earlier tokens, which provide critical context, and gradually propagate to later tokens. This iterative re-masking and refinement process allows MMHMR to systematically resolve ambiguities, enhance prediction accuracy, and generate robust, high-fidelity 3D hand mesh reconstructions.

B.6. Token sampling strategies

The results demonstrate that the Top-k sampling strategy has a significant impact on MMHMR's performance as depicted in Table 4. Specifically, using a Top-1 sampling approach yields the best results, with the lowest PA-MPJPE and MVE across both the HO3Dv3 and FreiHAND datasets (Table 4). As the value of k increases, there is a noticeable decline in performance, with higher k values leading to increased PA-MPJPE and MVE, indicating reduced accuracy in 3D pose estimation and mesh reconstruction. This suggests that restricting the model to the most confident token predictions (Top-1) is crucial for maintaining high precision in MMHMR, while broader sampling (higher k) introduces noise and uncertainty that degrade the overall performance of the estimated final pose.

B.7. Impact of Reference Keypoints in Deformable Attention

B.8. Impact of Control Signals

B.9. Impact of Pose Tokenizer on MMHMR

The results from our experiments highlight the critical role of the Pose Tokenizer's design in the overall performance of MMHMR as shown in Table 6 and 7. Specifically, our ablation studies on the codebook size demonstrate that an increase in codebook size initially enhances performance, as seen when moving from a 1024×256 to a 2048×256 configuration, with notable improvements in both PA-MPJPE and MVE metrics across the HO3Dv3 and FreiHAND datasets (Table 6). However, further expansion to a 4096×256 codebook leads to a decline in accuracy, indicating that while larger codebooks can provide richer pose representations, they may also introduce complexity that hinders 3D pose estimation in subsequent stages. Moreover, the number of tokens in the second stage critically impacts MMHMR's performance (Table 7). A moderate token count (96) provides the best results, with lower (48) and higher (192, 384) counts leading to reduced accuracy. This highlights the need for an optimal balance between pose representation

Table 3. Impact of different loss combinations on PA-MPJPE and PA-MPVPE errors for the HO3Dv3 and FreiHAND datasets. Results are based on the UGS stage with 5 iterations, using initial 3D pose estimates.

Used Losses	**HO3Dv3**		**FreiHAND**	
	PA-MPJPE ↓	PA-MPVPE ↓	PA-MPJPE ↓	PA-MPVPE ↓
$L_{mask}, L_\theta, L_{3D}, L_{2D}, \beta$	42.1	68.1	77.5	51.7
$L_{mask}, L_\theta, L_{3D}, \beta$	43.7	78.5	90.0	56.9
$L_{mask}, L_\theta, \beta$	45.1	80.0	91.5	58.9
L_{mask}, β	45.5	80.5	92.5	60.8
L_θ, β	48.5	82.0	95.6	61.7
L_{3D}, β	57.9	87.5	146.4	67.9
L_{2D}, β	103.6	1160.6	1167.7	110.7

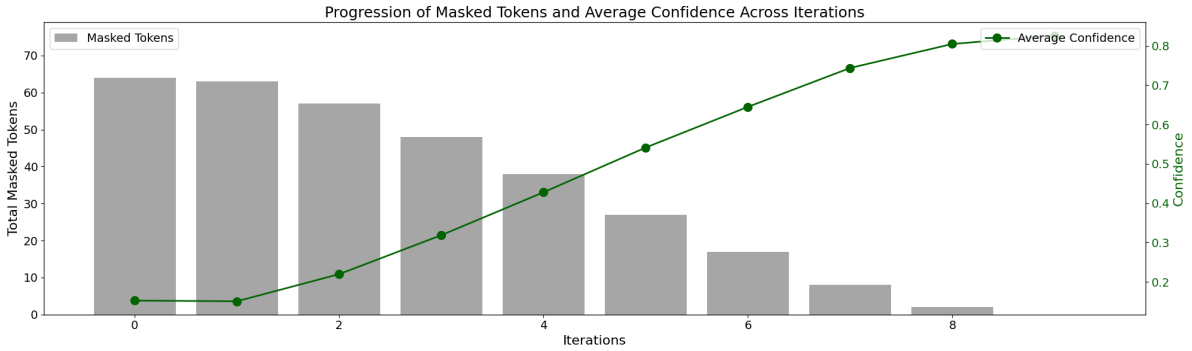

Figure 2. This figure illustrates the progression of masked tokens across iterations. **Legend:** ■ represents masked ([MASK]) tokens, shown in dark green, while ■ indicates unmasked tokens, shown in light cyan. The horizontal axis represents token indices, while the vertical axis corresponds to iterations. Darker colors emphasize masked areas, while lighter colors signify the gradual unmasking of tokens as iterations progress. This visualization highlights the dynamics of the masking schedule over time. Heatmap visualization of the Uncertainty-Guided Sampling Process. The heatmap illustrates the iterative decoding of a masked sequence over T iterations. The color gradient reflects prediction confidence, with [fill=yellow,draw=black] (0,0) rectangle (0.2,0.2); representing high confidence and [fill=customblue,draw=black] (0,0) rectangle (0.2,0.2); representing low confidence.

Table 4. Impact of Top-k Sampling on MMHMR. Here, results are from **5 UGS iterations**, refining pose tokens to evaluate initial 3D pose estimates at inference.

	HO3Dv3		FreiHAND	
Top-K	MPJPE	MVE	MPJPE)	MVE
1	**68.1**	**77.5**	**88.2**	**99.5**
2	7.0	7.0	98.2	116.9
5	85.1	98.6	108.6	123.1
10	90.5	110.1	118.1	130.6
20	98.9	120.6	130.9	140.3

Table 5. Mesh-Guided Control for Hand Image Generation

HaMer Mesh Estimator	**FID-H ↓**	**KID-H ↓**	**Hand Det Conf. ↑**
2D Pose Guidance [46]	45.65	0.035	0.865
Mesh Guidance	40.23	0.027	0.912
Mesh+2D Pose Guidance	40.23	0.027	0.912

richness and the model's ability to effectively utilize this information for accurate 3D hand mesh reconstruction.

Table 6. Impact of Codebook Size (Tokens = 96) on MMHMR. Here, results are from **5 iterations of UGS**, iteratively refining pose tokens to evaluate initial 3D pose estimates at inference

	HO3Dv3		FreiHAND	
# of code × code dimension	MPJPE (↓)	MVE (↓)	MPJPE (↓)	MVE (↓)
1024 × 256	70.1	82.5	90.2	103.7
2048 × 128	69.9	80.3	89.2	99.9
2048 × 256	**68.1**	**77.5**	**88.2**	**99.5**
4096 × 256	69.5	80.4	90.4	100.4

B.10. Ablation of Feature Resolutions

The ablation study on feature resolutions in MMHMR provides crucial insights into the efficacy of multi-scale feature representation for hand Mesh Recovery. Results, as shown in Table **??**, consistently demonstrate that increasing resolution from 1× to 16× significantly enhances accuracy, reducing PA-MPJPE by 3.2 mm on HO3Dv3 and

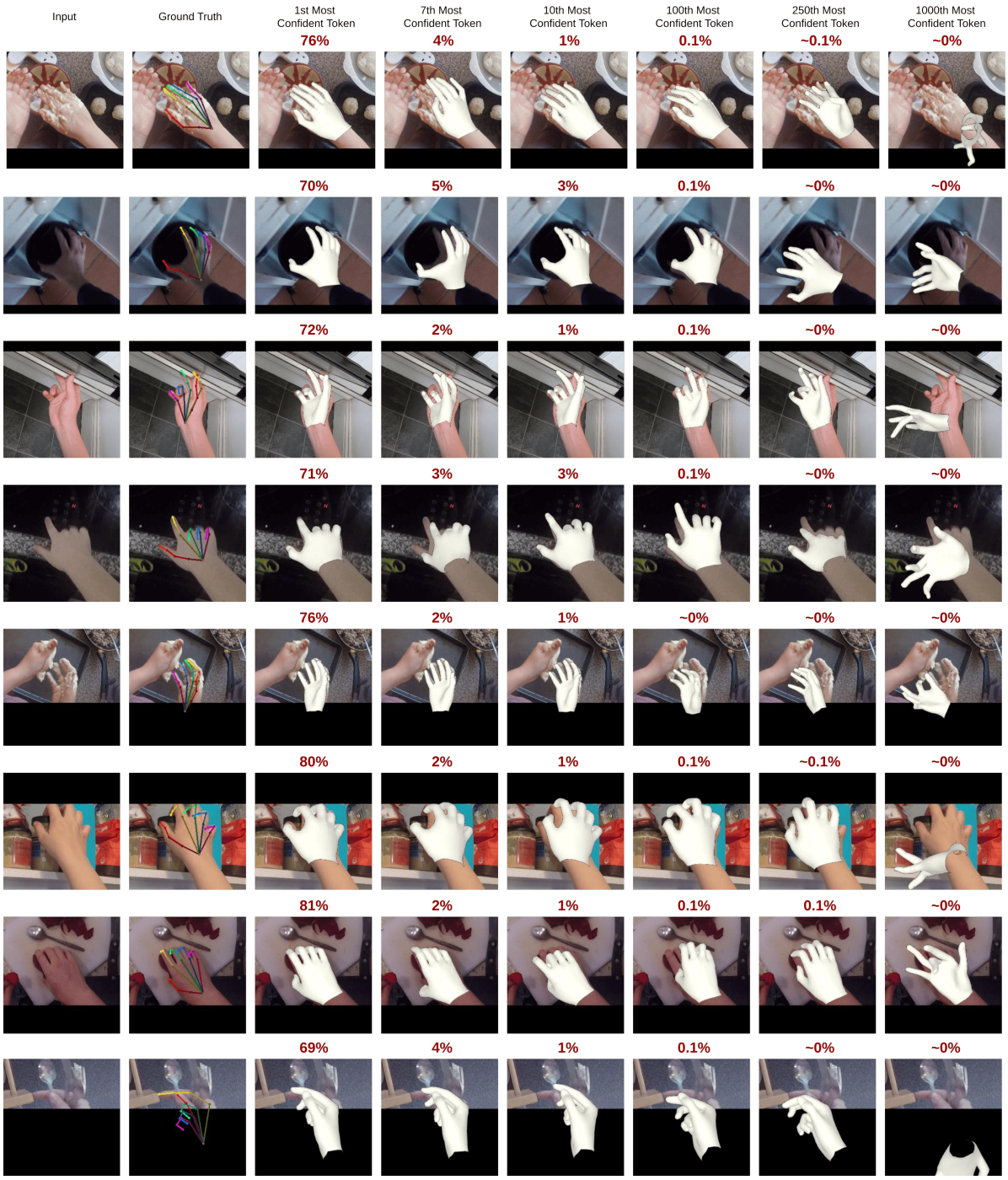

Figure 3. State-of-the-art (SOTA) methods, such as HMR2.0 [**?**] and TokenHMR [18], utilize vision transformers to recover 3D hand meshes from single images. However, the limitations of these SOTA approaches, particularly in dealing with unusual poses or ambiguous situations, are evident in the errors marked by red circles. Our approach, MMHMR, addresses these challenges by explicitly modeling and mitigating uncertainties in the 2D-to-3D mapping process, leading to more accurate and robust 3D pose reconstructions in complex scenarios.

4.1 mm on FreiHAND. However, further increases to 32× yield diminishing returns (0.2 mm on HO3Dv3, 0.1 mm on FreiHAND), indicating an optimal balance between performance and computational efficiency at 16× resolution. Cru-

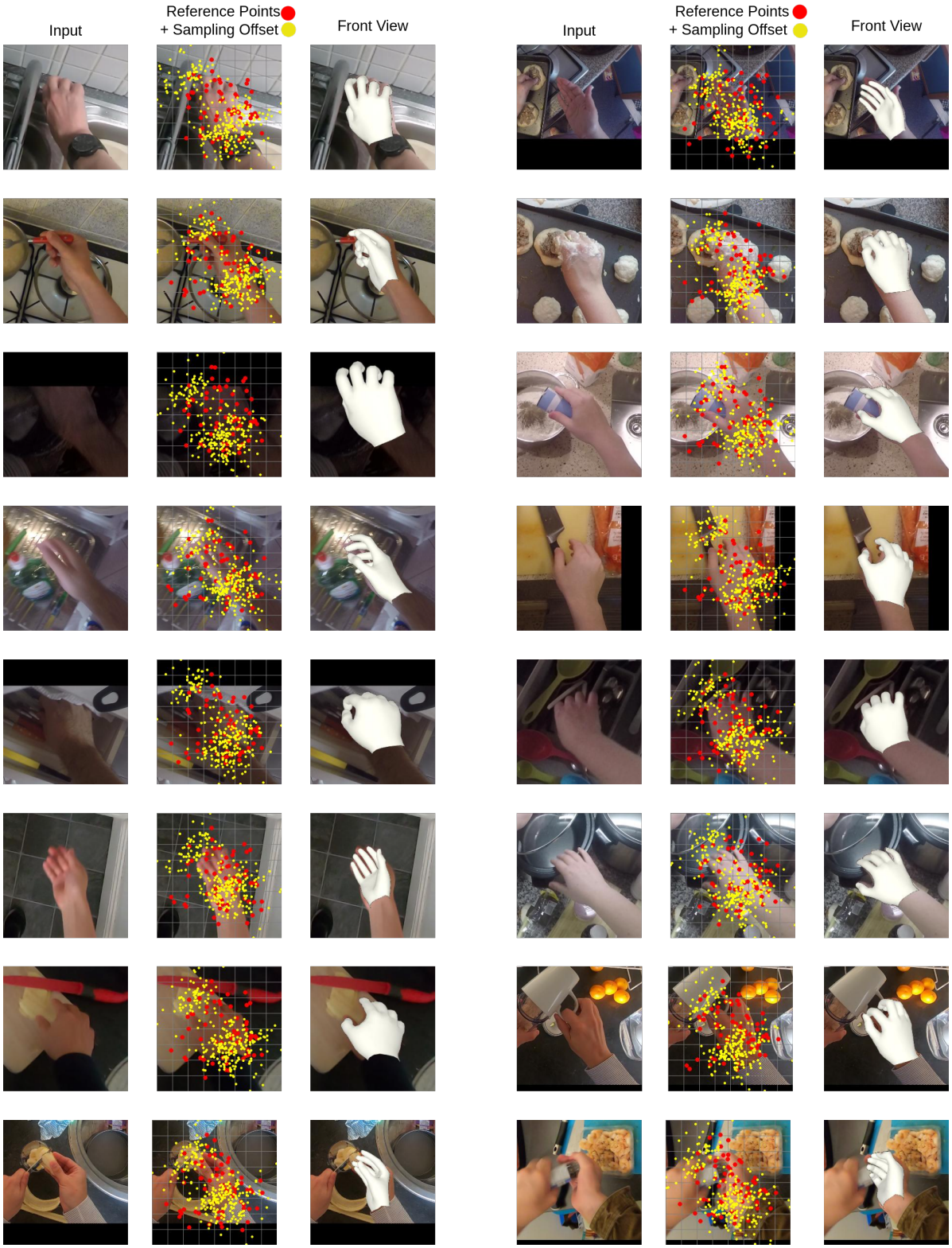

Figure 4. Impact of Reference Keypoints in Deformable Attention

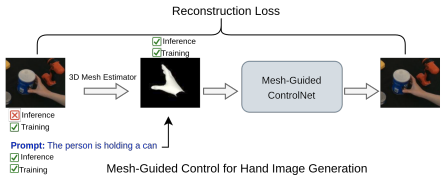

Figure 5. Generation Pipeline: MMHMR extracts 3D hand meshes from images during training, using them with text prompts to guide ControlNet. At inference, only text and mesh signals are needed to generate realistic hand images.

Table 7. Impact of Tokens (Codebook = 2048x256) on MMHMR.Here, results are from **5 iterations of UGS**, iteratively refining pose tokens to evaluate initial 3D pose estimates at inference

	HO3Dv3		FreiHAND	
# of code × code dimension	MPJPE	MVE	MPJPE	MVE
48	68.5	80.5	92.6	105.6
96	**68.1**	**77.5**	**88.2**	**99.5**
192	72.5	83.6	96.8	110.7
384	82.5	91.5	101.4	135.4

cially, the study reveals a symbiotic relationship between high and low-resolution features, with performance degrading when lower scales (1× or 4×) are omitted. This underscores the importance of MMHMR's multi-scale approach in capturing both fine-grained details and overall structure.

Table 8. Impact of feature resolutions on PA-MPJPE and PA-MPVPE errors for HO3Dv3 and FreiHAND datasets. Results are from 5 iterations of UGS, iteratively refining pose tokens to evaluate the initial pose estimate at inference.

Feature Scales (Included)	**HO3Dv3 ↓**		**FreiHAND ↓**	
	PA-MPJPE	**PA-MPVPE**	**PA-MPJPE**	**PA-MPVPE**
4×, 8×, 16×, 32×	**68.3**	**88.1**	**88.3**	**92.1**
4×, 8×, 16×	68.1	88.2	88.2	92.2
4×, 8×, 32×	68.6	89.0	89.1	92.9
4×, 8×	68.8	89.3	89.3	93.0
8×, 16×, 32×	69.3	90.0	89.8	93.5
4×	69.8	90.5	90.3	94.0
4×, 16×, 32×	69.8	90.3	89.9	93.7

B.11. Impact of Deformable Cross Attention Layers

The results show that the number of Deformable Cross Attention Layers is crucial for MMHMR's performance. Increasing the layers from 2 to 4 significantly improves PA-MPJPE and MVE on the AMASS and MOYO datasets, enhancing 3D pose estimation and mesh reconstruction as shown in Table 9. However, adding more than 4 layers leads to diminishing returns and slight performance degradation. This indicates that 4 layers provide the optimal balance between complexity and performance, ensuring MMHMR achieves accurate and efficient 3D hand mesh reconstruction.

Table 9. Impact of # of Deformable Cross Attention Layers in MMHMR. Here, results are from **5 iterations of UGS**, iteratively refining pose tokens to evaluate initial 3D pose estimates at inference

	HO3Dv3		FreiHAND	
# of Deformable Cross Attention Layers	MPJPE	MVE	MPJPE	MVE
2	70.5	85.9	94.8	107.0
4	68.1	77.5	88.2	**99.5**
6	69.5	**77.1**	**88.1**	100.2
8	70.9	79.9	91.9	104.9

B.12. Qualitative Results

We present qualitative results of MMHMR in Figures 9, 8, and **??**, demonstrating the model's robustness in handling extreme poses and partial occlusions. These results highlight the effectiveness of our approach, where reconstructions are well-aligned with the input images and remain valid when viewed from novel perspectives. A key factor contributing to this success is MMHMR's explicit modeling and reduction of uncertainty during the 2D-to-3D mapping process. By iteratively refining pose estimates and focusing on high-confidence predictions, MMHMR is able to mitigate the challenges that typically hinder other state-of-the-art methods. This approach ensures more accurate and consistent 3D reconstructions, even in complex scenarios where traditional deterministic models often falter. Additionally, the refinement process, as illustrated in Figure **??**, plays a crucial role in aligning 3D outputs with 2D pose detections, further enhancing the model's ability to produce realistic and accurate meshes.

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 3

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2

[4] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using

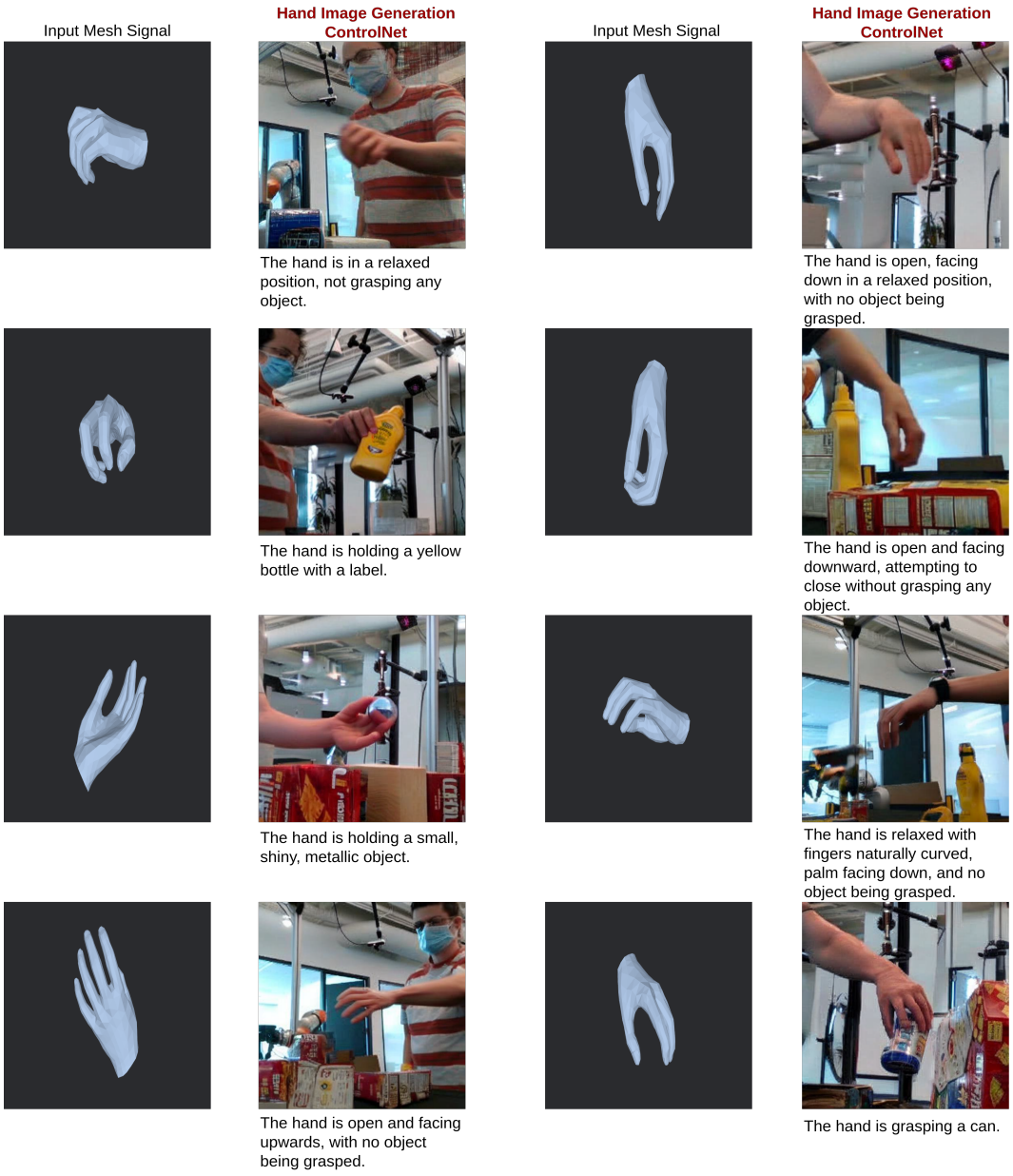

Figure 6. Control Signals

discriminative salient points. In *European Conference on Computer Vision (ECCV)*, 2012. 1

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019. 2

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 2, 5

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[8] 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. 5, 7, 9

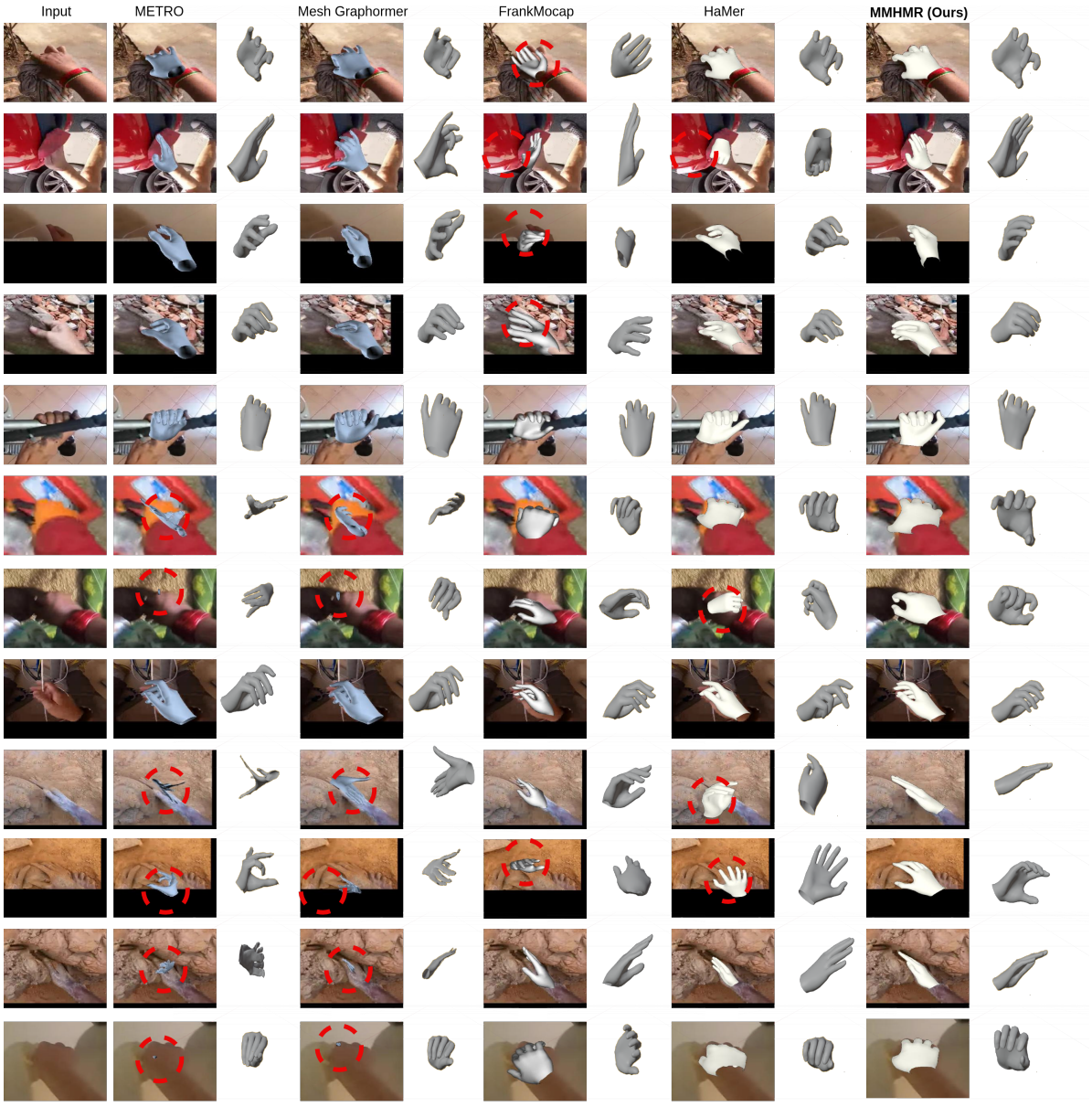

Figure 7. Comparison of State-of-the-Art (SOTA) Methods: Our method, MMHMR, synthesizes unobserved parts through generative modeling, enabling accurate 3D reconstructions in complex, occluded scenarios.

[9] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[10] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10451–10460, 2021. 6

[11] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. *Advances in Neural Information Processing Systems*, 36:30453–30465, 2023. 7

[12] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[13] Youwang Kim Oh Tae-Hyun Cho, Junhyeong. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pages 342–359. Springer, 2022. 6

Figure 8. Qualitative results of our approach on challenging poses from the LSP [33] dataset.

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 7

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv. *arXiv preprint arXiv:1810.04805*, 2019. 2

[16] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 2

[17] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Com-*

CVPR
#5785

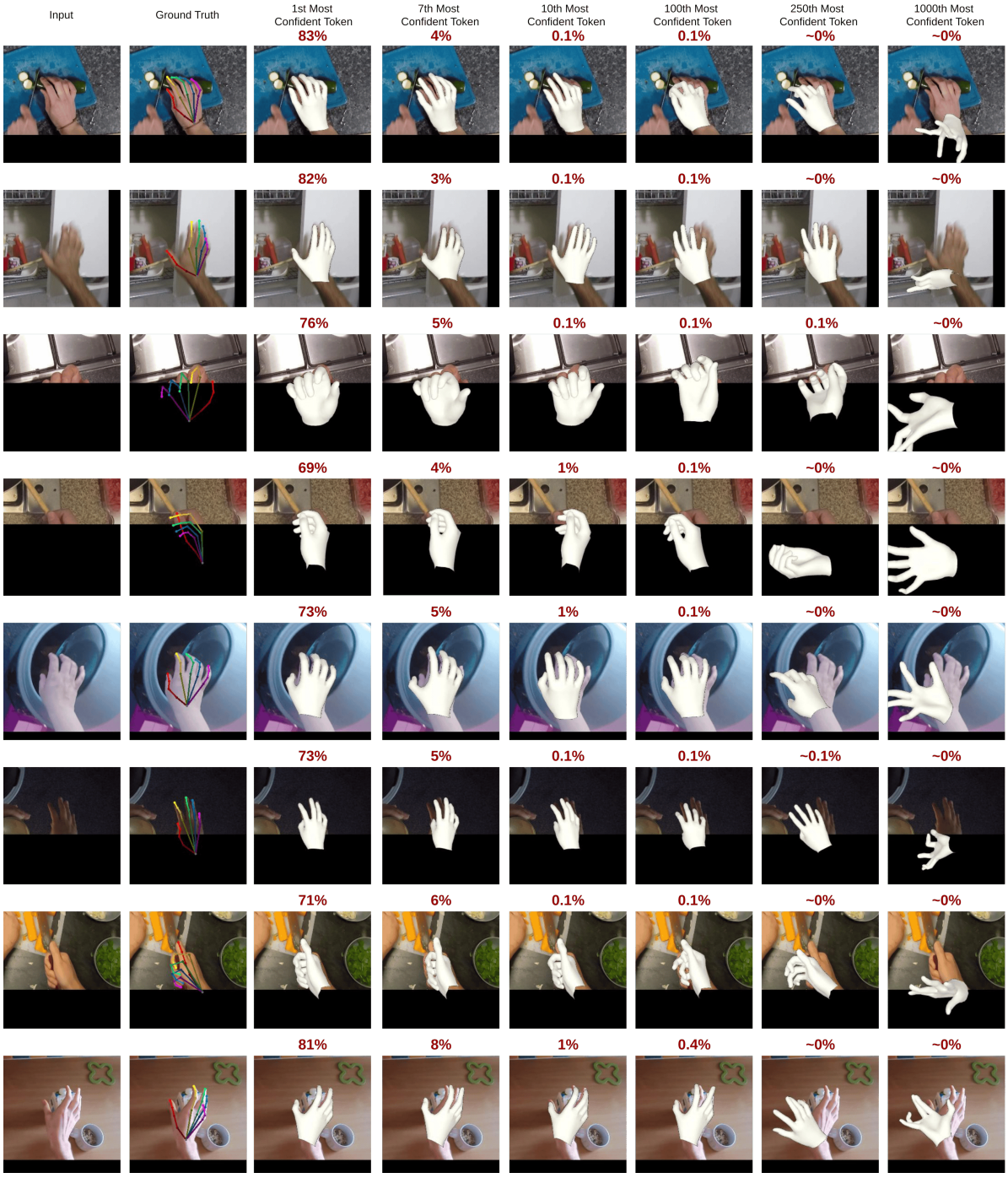

Figure 9. State-of-the-art (SOTA) methods, such as HMR2.0 [**?**] and TokenHMR [18], utilize vision transformers to recover 3D hand meshes from single images. However, the limitations of these SOTA approaches, particularly in dealing with unusual poses or ambiguous situations, are evident in the errors marked by red circles. Our approach, MMHMR, addresses these challenges by explicitly modeling and mitigating uncertainties in the 2D-to-3D mapping process, leading to more accurate and robust 3D pose reconstructions in complex scenarios.

puter Vision, pages 6353–6363, 2024. 6

[18] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024. 13, 19

Figure 10

[19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[20] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022. 5

[21] Chengying Gao, Yujia Yang, and Wensheng Li. 3d interacting hand pose and shape estimation from a single rgb image. *Neurocomputing*, 2022. 2

[22] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[23] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *arXiv preprint arXiv:1707.03742*, 2017. 1

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 7

[25] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 5

[26] 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. 6, 7, 8

[27] 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. 9

[28] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: perception of finger motion from reduced marker sets. In *SIGGRAPH*, 2012. 2

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[30] Rongtian Huo, Qing Gao, Jing Qi, and Zhaojie Ju. 3d human pose estimation in video for human-computer/robot interaction. In *Intelligent Robotics and Applications*, pages 176–187, Singapore, 2023. Springer Nature Singapore. 1

[31] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 758–767, 2023. 6

[32] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 5

[33] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011. 18

[34] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is matter: Point-guided 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12880–12889, 2023. 6

[35] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326*, 2019. 2

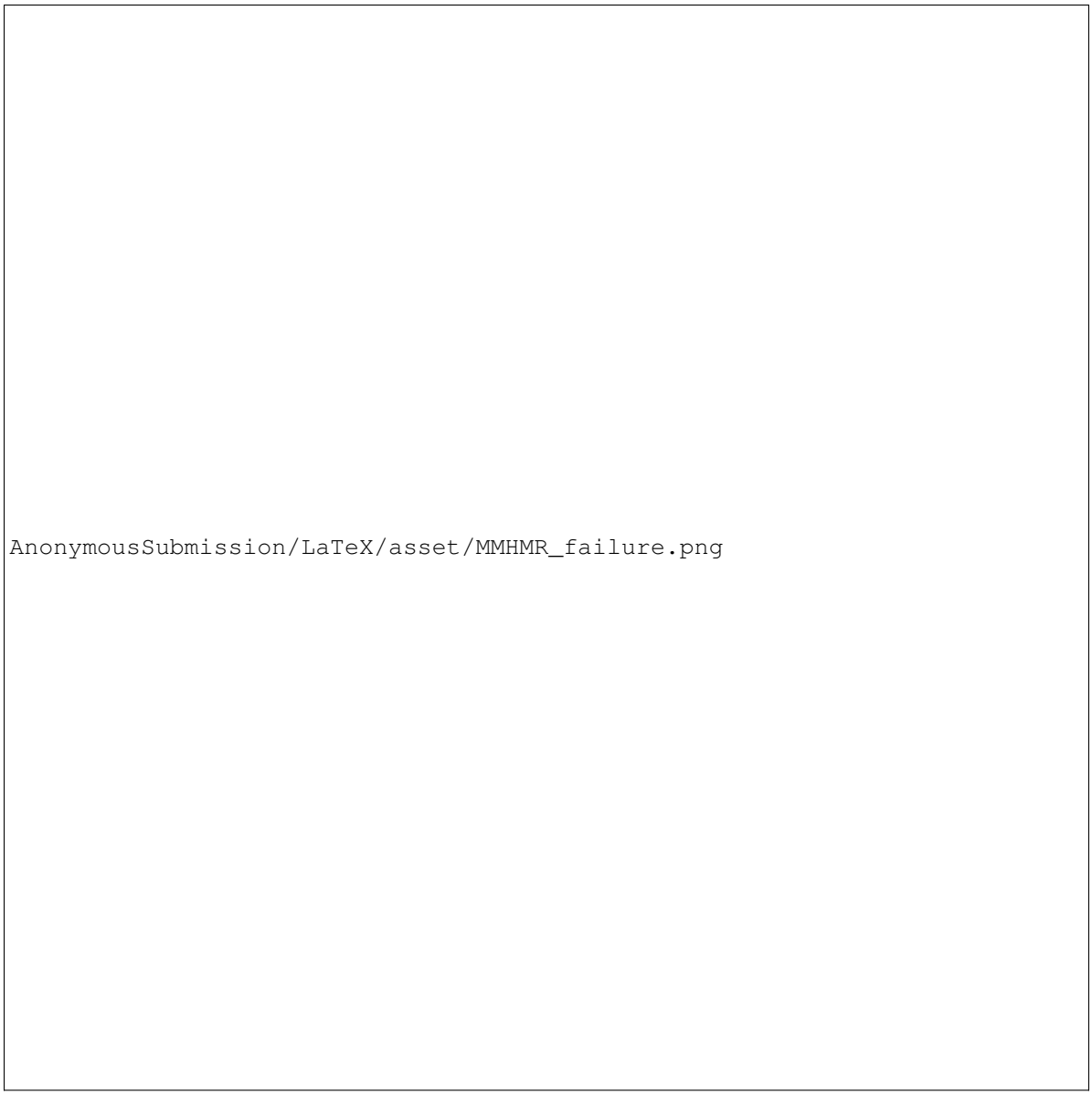

Figure 11. Failure Cases of MMHMR in 3D hand Reconstruction: MMHMR often encounters errors when dealing with unusual body articulations and complex depth ordering of body parts. These challenges typically result in inaccurate 3D poses and non-valid outputs. The root of this limitation lies in the model's reliance on the MANO parametric model, which may not fully capture the complexity of extreme or uncommon hand poses.

[36] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2

[37] 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. 2, 5, 6

[38] 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. 7

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 1, 2, 6, 7

[40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan

Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 9

[41] Haofan Lu, Shuiping Gou, and Ruimin Li. Spmhand: Segmentation-guided progressive multi-path 3d hand pose and shape estimation. *IEEE Transactions on Multimedia*, 2024. 6

[42] Anders Markussen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1073–1082, New York, NY, USA, 2014. Association for Computing Machinery. 1

[43] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[44] 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. 5, 9

[45] 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. 7

[46] 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. 1, 2, 3, 5, 6, 7, 12

[47] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[48] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3

[49] 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. 7

[50] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 1, 5

[51] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 1

[52] Srinath Sridhar, Anna Maria Feit, Christian Theobalt, and Antti Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 3643–3652, New York, NY, USA, 2015. Association for Computing Machinery. 1

[53] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[54] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *TOG*, 2017. 2

[55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[56] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023. 6

[57] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *TOG*, 2020. 2

[58] Shuaibing Wang, Shunli Wang, Dingkang Yang, Mingcheng Li, Ziyun Qian, Liuzhen Su, and Lihua Zhang. Handgcat: Occlusion-robust 3d hand mesh reconstruction from monocular images. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2495–2500. IEEE, 2023. 6

[59] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020. 3

[60] 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. 5, 9

[61] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2750–2760, 2022. 6

[62] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2023. 6

[63] Ziwei Yu, Linlin Yang, You Xie, Ping Chen, and Angela Yao. Uv-based 3d hand-object reconstruction with grasp optimization. *arXiv preprint arXiv:2211.13429*, 2022. 6

[64] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 5

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023. 7, 9

[66] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 2

[67] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *ICCV*, 2021. 2

[68] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. Ufc-bert: Unifying multi-modal controls for conditional image synthesis. *Advances in Neural Information Processing Systems*, 34:27196–27208, 2021. 2

[69] Wenping Zhao, Jinxiang Chai, and Ying-Qing Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *SIGGRAPH*, 2012. 2

[70] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2024. 5, 6

[71] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020. 4

[72] 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. 5, 9

[73] 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. 5, 6, 7, 8, 10