# EXDIFF:DIFFUSION AS AN EXPRESSIVE PRIOR FOR HAND GESTURES AND FACIAL EXPRESSIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Expressive hand gestures and facial expressions provide essential non-verbal cues for understanding human intent in image-based human-computer interaction. However, existing pose priors often focus on full-body motion or model the hands and face as independent components, limiting their ability to capture the dependencies between hand articulation and facial expression. We propose EXDIFF, an expressive diffusion framework for jointly modeling bilateral hand gestures, jaw motion, and facial expressions within a unified expressive latent space. The proposed framework introduces part-balanced diffusion with part-aware noise schedules, cross-part residual fusion, and geometry-aware denoising supervision to model fine-grained hand articulation, facial dynamics, and their mutual dependencies. To support heterogeneous training, EXDIFF learns part-specific hand and face diffusion priors, then trains a zero-initialized cross-part residual fusion head on joint hand–face data, using part-availability masks and confidence-aware conditioning. The learned prior supports expressive generation, missing-part completion, and monocular hand–face mesh recovery. Experiments show that EXDIFF consistently improves realism, coverage, and diversity on ARCTIC, FreiHAND, NOW, and BEAT2. The code will be made available to the public.

## 1 INTRODUCTION

Human communication is fundamentally expressed through fine-grained hand gestures and facial expressions. Hands convey interaction, intention, object manipulation, and symbolic communication, while facial expressions provide important cues about emotion, attention, and social response. Accurate modeling of these expressive regions is therefore essential for human-computer interaction, sign language understanding, embodied intelligence, and related applications. However, recovering and generating coordinated hand articulation and facial expression from monocular images remains highly challenging due to ambiguous viewpoints, rapid finger articulation, subtle jaw and facial expression changes, and the limited availability of datasets that jointly annotate both hands and face.

Existing pose priors treat full-body pose, hand articulation, and facial expression as separate problems. Full-body priors such as GMM Bogo et al. (2016), VPoser Pavlakos et al. (2019), NRDF He et al. (2024), and DPoser-X Lu et al. (2025) provide global constraints but spend most of their capacity on torso and limb parameters that are irrelevant to expressive hand–face modeling, whereas hand-specific Baek et al. (2020); Cao et al. (2021); Chen et al. (2021b); Liu et al. (2021); Yang et al. (2021b); Zhao et al. (2021); Pavlakos et al. (2024); Dong et al. (2024); Saleem et al. (2025) and face-specific Chai et al. (2023); Daněček et al. (2022; 2023); Wood et al. (2022); Lei et al. (2023) methods capture local geometry in detail but ignore the coordination between hand gestures and facial expressions. Expressive behavior in human–computer interaction is inherently cross-part, involving synchronized hand motion, mouth movement, gaze, and conversational intent. We address one well-defined component of this problem, the statistical dependency between finger articulation, jaw motion, and facial expression in part-local frames; our claims are scoped accordingly. We therefore propose EXDIFF, an expressive diffusion prior over the left hand, right hand, and face only. Using SMPL-X Pavlakos et al. (2019) solely as a geometric scaffold for fitting and decoding, EXDIFF excludes body pose from the diffusion representation, reducing complexity while emphasizing fine-grained finger, jaw, and facial dynamics. As shown in Fig. 1, a single model supports monocular expressive recovery, joint hand–face generation, and diverse hand-gesture and facial-expression generation.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

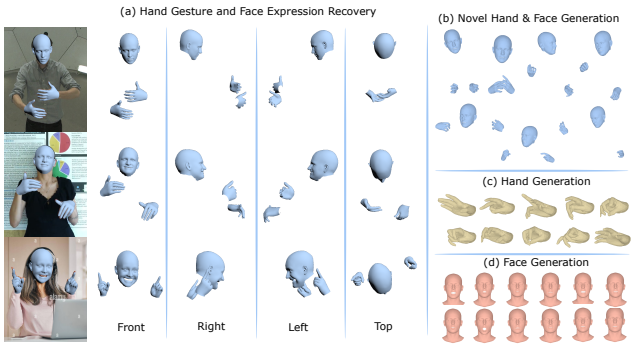


Figure 1: Qualitative overview of expressive hand–face modeling. EXDIFF reconstructs coordinated hand gestures and facial expressions from monocular images and visualizes the recovered geometry from multiple views (a). It also generates diverse joint hand–face configurations (b), plausible hand articulations (c), and varied facial expressions (d).

The core of EXDIFF is a part-balanced score-based diffusion model trained on a unified expressive representation of the face and hand, consisting of the left-hand pose, right-hand pose, jaw pose and facial expression coefficients. To accommodate the different dimensionalities and motion statistics of hand articulation and facial expressions, EXDIFF introduces region-adaptive diffusion schedules, part-balanced score matching, and part-availability masks. This formulation enables stable training within a single diffusion space using heterogeneous supervision from hand-only, face-only, and joint hand–face samples. The proposed part-aware denoising network combines shared hand priors, a dedicated face branch, and cross-part residual fusion to model dependencies between bilateral hand gestures and facial expressions while preserving stable part-specific predictions. To further enforce geometric plausibility, EXDIFF incorporates geometry-aware denoising supervision by decoding the predicted clean expressive parameters through SMPL-X Pavlakos et al. (2019) and supervising the resulting 3D hand joints and facial landmarks. During inference, confidence-aware guidance leverages reliable observations for reconstruction, while the learned diffusion prior completes missing or uncertain expressive components. To the best of our knowledge, EXDIFF is the first diffusion-based pose prior focused exclusively on bi-hands gesture and face expression. This design choice reduces computational complexities associated with full-body diffusion models while improving the modeling capacity for fine-grained expressive behavior.

The main contributions are summarized as follows:

- We propose EXDIFF, a novel expressive diffusion framework focusing on coordinated hand gestures and facial expressions for hand–face mesh recovery and generation.
- We develop a part-balanced expressive diffusion formulation and region-adaptive noise schedules, allowing stable training across hand and face components.
- We design a part-aware denoising network with part-specific priors and cross-part residual fusion, enabling the model to capture dependencies between bilateral hand gestures and facial expressions.
- Extensive qualitative and quantitative results demonstrate that EXDIFF improves realism, fidelity, coverage, and diversity compared to previous models.
- Extensive qualitative and quantitative results demonstrate that EXDIFF improves realism, fidelity, coverage, and diversity compared to previous state-of-the-art models.

## 2 RELATED WORK

**Hand and Face Mesh Recovery.** Hand mesh recovery is dominated by parametric methods that regress MANO Romero et al. (2017) pose and shape from monocular images Hasson et al. (2019); Baek et al. (2020); Boukhayma et al. (2019); Cao et al. (2021); Chen et al. (2021b); Jiang et al. (2021); Liu et al. (2021); Yang et al. (2020; 2021b); Zhang et al. (2021a; 2019; 2021b); Pavlakos et al. (2024); Dong et al. (2024); Saleem et al. (2025), complemented by voxel- and heatmap-based methods with stronger spatial supervision Boukhayma et al. (2019); Iqbal et al. (2018); Moon & Lee (2020); Moon et al. (2020); Yang et al. (2021a) and by direct vertex regression Lin et al. (2021a); Chen et al. (2021a; 2022); Ge et al. (2019); Kulon et al. (2020). Face modeling likewise relies on parametric models such as FLAME Li et al. (2017) and the face component of SMPL-X Pavlakos et al. (2019), which compactly encode jaw motion, shape, and expression. Despite strong accuracy,

both families are designed for a single region and do not model the dependence between hand articulation and facial expression. EXDIFF instead models the left hand, right hand, and face within a unified generative representation.

**Diffusion Priors for Expression and Gesture.** DPoser Lu et al. (2025) learns a score-based prior over SMPL-X parameters as a plug-in fitting regularizer, and extends it to the whole body with part priors and masked training on mixed data, while NRDF He et al. (2024) learns a Riemannian distance field that avoids Euclidean noise on rotations. Following DPoser-X, EXDIFF restricts it to the expressive parts and alters the training signal through a dimension-normalized score loss, a zero-initialized residual head trained on paired data with frozen part priors, and geometry supervision after SMPL-X decoding. Speech-driven generators such as DiffSHEG Chen et al. (2024) and EMAGE Liu et al. (2024), which introduces BEAT2, model temporal expression–gesture dependencies conditioned on audio, whereas EXDIFF is a static, observation-agnostic prior for monocular fitting; the two are complementary, and BEAT2 serves as one of our evaluation sets.

**Difference from Prior Work.** Unlike DiffMesh Zheng et al. (2025) and DPoser-X Lu et al. (2025), which learn whole-body diffusion priors, EXDIFF models only the left hand, right hand, and face, avoiding unnecessary body parameters. EXDIFF adds (i) part-specific schedules, (ii) dimension-normalized part balancing, (iii) a zero-initialized residual fusion head trained only on paired hand–face data, and (iv) geometry-aware supervision; each is ablated in. Unlike hand-only Dong et al. (2024); Kim et al. (2023); Pavlakos et al. (2024); Zhou et al. (2024) or face-only priors Daněček et al. (2022); Feng et al. (2021); Retsinas et al. (2024); Zhang et al. (2023), EXDIFF learns a unified hand–face representation from heterogeneous hand-only, face-only, and paired data, and its part-aware denoiser with shared hand priors, a dedicated face branch, and cross-part residual fusion captures dependencies between bilateral gestures and facial expression, yielding a single prior for generation and monocular fitting.

# 3 METHODOLOGY

## 3.1 OVERVIEW

We propose EXDIFF, an expressive diffusion framework for coordinated hand–face modeling from monocular images; video is processed frame by frame, and EXDIFF contains no temporal component. Instead of modeling the full body, EXDIFF focuses on the left hand, right hand, and face, where each expressive part follows a part-specific noise schedule Fig. 2(a). The model first forms a unified expressive representation and processes it through a part-aware denoising network Fig. 2(b), equipped with cross-part residual fusion block Fig. 2(b.1) conditioned on timestep, part mask, and observation confidence, and a zero-initialized residual head Fig. 2(b.2) refines the shared hand and face priors Fig. 2(b.3)(b.4), to predict clean samples Fig. 2(b.5). SMPL-X Pavlakos et al. (2019) is used only as a geometric scaffold for the fitting, decoding, and visualization of multiple views Fig.2 (c), while the torso, leg, and full-body pose parameters are excluded from the diffusion representation. Concretely, hand poses are expressed in wrist-local coordinate frames, and the jaw pose in a head-local frame; during decoding and geometry-aware supervision, the SMPL-X body is fixed to the neutral pose with identity global orientation, so that hand joints and facial landmarks are evaluated in a canonical body-anchored space. Global orientation and camera parameters are handled exclusively by the fitting stage and are never part of the diffusion representation.

## 3.2 FACE-HAND EXPRESSIVE REPRESENTATION

Let $\mathcal{P} = \{lh, rh, face\}$ denote the set of expressive parts corresponding to the left hand, right hand, and face. The expressive parameter vector is defined as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{lh} \\ \mathbf{x}^{rh} \\ \mathbf{x}^{face} \end{bmatrix}, \qquad \mathbf{x}^{face} = \begin{bmatrix} \mathbf{x}^{jaw} \\ \mathbf{x}^{expr} \end{bmatrix}, \tag{1}$$

Here, $\mathbf{x}^{lh}$ and $\mathbf{x}^{rh}$ denote the left- and right-hand pose parameters, $\mathbf{x}^{jaw}$ denotes the jaw pose parameters, and $\mathbf{x}^{expr}$ denotes the facial expression coefficients. The torso, leg, and full-body pose parameters are excluded from the diffusion representation.

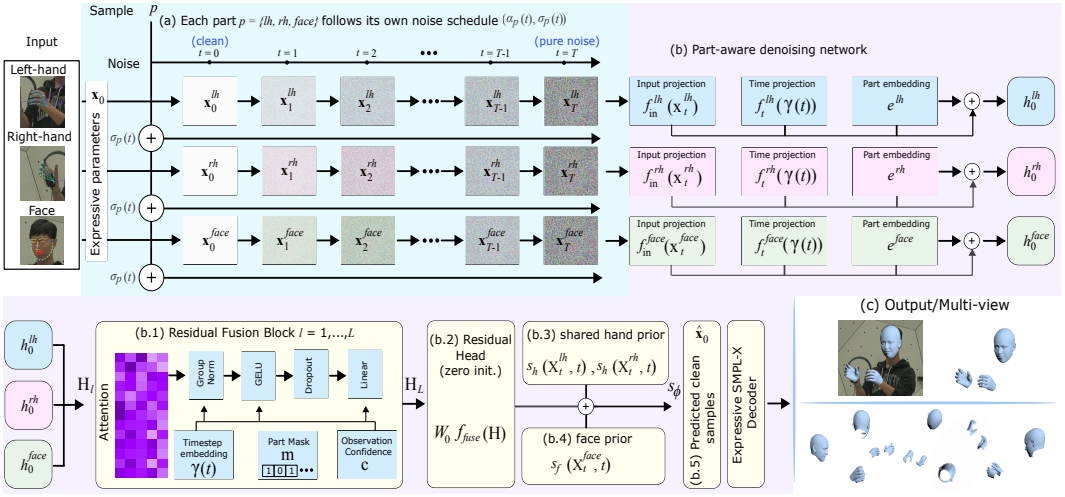


Figure 2: Illustration of EXDIFF framework. Each expressive part follows its own variance-exploding noise schedule; part embeddings are fused by $L$ cross-part residual attention blocks conditioned on the timestep, part mask $\mathbf{m}$, and observation confidence $\mathbf{c}$, and a zero-initialized residual head refines the shared hand and face priors to predict clean samples.

The dimensionality of the expressive vector is

$$D = 2N_h d_r + N_j d_r + N_e, \tag{2}$$

where $N_h$ is the number of joints per hand, $N_j$ the number of jaw joints, $N_e$ the number of expression coefficients, and $d_r$ the rotation dimension. All experiments use axis-angle rotations ($d_r = 3$). Although Gaussian noise on axis-angle vectors ignores the geometry of $SO(3)$, the finger and jaw rotations in our training data are anatomically small, with norms well below $\pi$, so the normalized vectors stay far from the antipodal discontinuity; a continuous 6D representation Zhou et al. (2019) ($d_r = 6$) is compatible with the formulation and is left for future work. Each component is normalized by a shared face–hand normalizer with per-dimension mean and standard deviation computed over the combined training corpus, and $\mathbf{x}$ denotes the normalized vector hereafter. Eq. equation 1 covers articulation and expression only; for face-shape generation, FLAME Li et al. (2017) shape coefficients $\boldsymbol{\beta}$ are modeled by the face branch alongside $\mathbf{x}^{expr}$ (Sec. 4.1).

### 3.3 PART-BALANCED EXPRESSIVE DIFFUSION

The diffusion prior is trained over the normalized expressive vector $\mathbf{x}$. Since hand articulation and facial expression have different dimensions and motion statistics, EXDIFF adopts a part-specific forward diffusion process. For each expressive part $p \in \mathcal{P} = \{lh, rh, face\}$, a clean sample $\mathbf{x}_0^p$ is perturbed as

$$\mathbf{x}_t^p = \alpha_p(t)\mathbf{x}_0^p + \sigma_p(t)\boldsymbol{\epsilon}^p, \qquad \boldsymbol{\epsilon}^p \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $t \in [0, 1]$ is the continuous diffusion time, discretized into $T = 1000$ noise scales, and $\alpha_p(t)$, $\sigma_p(t)$ are the signal and noise schedules of part $p$. Noise is drawn independently per part, so the parts sit at different relative noise levels at the same $t$. We instantiate the schedules with the variance-exploding formulation of Song et al. (2020):

$$\alpha_p(t) = 1, \qquad \sigma_p(t) = \sigma_{\min}^p \left( \frac{\sigma_{\max}^p}{\sigma_{\min}^p} \right)^t, \qquad t \in [0, 1], \tag{4}$$

so that $\sigma_p(t)$ grows geometrically from $\sigma_{\min}^p$ to $\sigma_{\max}^p$ without attenuating the signal. Both hands share one range, $(\sigma_{\min}^h, \sigma_{\max}^h) = (0.01, 50)$, since they are modeled by a shared hand prior, while the face uses a narrower range, $(\sigma_{\min}^{face}, \sigma_{\max}^{face}) = (0.005, 20)$. The wider hand range pushes the high-dimensional finger parameters fully into the noise regime, which is needed for diverse gesture sampling and for correcting large pose errors during fitting, whereas the narrower face range preserves the low-variance jaw and expression structure that the face prior must recover. The denoising

network estimates the conditional score function as

$$s_\phi(\mathbf{x}_t, t, \mathbf{m}, \mathbf{c}) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{m}, \mathbf{c}), \tag{5}$$

where $\mathbf{m} \in \{0,1\}^{|\mathcal{P}|}$ is the part-availability mask ($m_p = 1$ if part $p$ is annotated) and $\mathbf{c}$ denotes the observation confidence. These conditioning signals allow the model to handle complete, partial, and heterogeneous hand–face annotations.

To prevent the high-dimensional hand parameters from dominating training, we use a part-balanced score-matching objective:

$$\mathcal{L}_{\text{score}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon},\mathbf{m},\mathbf{c}} \left[ \sum_{p \in \mathcal{P}} \frac{\lambda_p \, m_p}{D_p} \left\| \sigma_p(t) \, s_\phi^p(\mathbf{x}_t, t, \mathbf{m}, \mathbf{c}) + \boldsymbol{\epsilon}^p \right\|_2^2 \right], \tag{6}$$

where $D_p$ is the dimensionality of part $p$, $\lambda_p$ controls the contribution of each region ($\lambda_p = 1$ for all $p$), and $m_p$ zeroes the loss of parts without valid supervision. Together with the part-specific schedules of Eq. 4, this objective balances the influence of each expressive component and enables unified training with mixed hand-only, face-only, and joint hand–face data.

### 3.4 Part-Aware Denoising Network

The denoising network combines part-specific priors with cross-part residual fusion. Each part is first embedded independently:

$$\mathbf{h}_0^p = f_{\text{in}}^p(\mathbf{x}_t^p) + f_t^p(\gamma(t)) + \mathbf{e}^p, \qquad p \in \mathcal{P}, \tag{7}$$

where $\gamma(t)$ is the timestep embedding and $\mathbf{e}^p$ is a learnable part embedding. The left and right hands share the same hand prior, while the face branch uses a separate face prior:

$$s_\phi^{base} = \left[ s_h(\mathbf{X}_t^{lh}, t), s_h(\mathbf{X}_t^{rh}, t), s_f(\mathbf{X}_t^{face}, t) \right]. \tag{8}$$

To model dependencies between hand gestures and facial expressions, the part features are fused using cross-part attention:

$$\mathbf{H}_0 = \left[ \mathbf{h}_0^{lh}, \mathbf{h}_0^{rh}, \mathbf{h}_0^{face} \right], \; \mathbf{H}_{l+1} = \mathbf{H}_l + g_l \left( \text{Attn}(\mathbf{H}_l), \gamma(t), \mathbf{m}, \mathbf{c} \right), \tag{9}$$

where $\text{Attn}(\cdot)$ denotes cross-part attention, and $g_l(\cdot)$ is a residual block with timestep conditioning, normalization, nonlinearity, and dropout.

The final score is obtained by adding a zero-initialized residual correction to the base part-wise prediction:

$$s_\phi(\mathbf{x}_t, t, \mathbf{m}, \mathbf{c}) = s_\phi^{base} + W_0 f_{\text{fuse}}(\mathbf{H}_L), \; W_0 = 0 \text{ at init.} \tag{10}$$

This design preserves the pretrained part-specific behavior at the beginning of finetuning and gradually learns cross-part dependencies between hand gestures and facial expressions.

### 3.5 Geometry-Aware Denoising Supervision

Parameter-space score matching alone does not guarantee geometrically plausible hand articulation or realistic facial motion. Therefore, ExDiff introduces geometry-aware denoising supervision on the predicted clean expressive sample. For each expressive part $p \in \mathcal{P}$, the part-wise clean estimate is recovered from the noisy input using the predicted score:

$$\hat{\mathbf{x}}_0^p = \frac{\mathbf{x}_t^p + \sigma_p^2(t) s_\phi^p(\mathbf{x}_t, t, \mathbf{m}, \mathbf{c})}{\alpha_p(t)}. \tag{11}$$

Here, $\alpha_p(t)$ and $\sigma_p(t)$ are the part-specific signal and noise schedules, $s_\phi^p(\cdot)$ is the predicted score for part $p$, $\mathbf{m}$ is the part-availability mask, and $\mathbf{c}$ denotes the confidence of observation.

The part-wise denoised estimates are then concatenated to form the complete denoised expressive vector:

$$\hat{\mathbf{x}}_0 = \begin{bmatrix} \hat{\mathbf{x}}_0^{lh} \\ \hat{\mathbf{x}}_0^{rh} \\ \hat{\mathbf{x}}_0^{face} \end{bmatrix}. \tag{12}$$

This complete vector is used for geometry decoding, ensuring that the left hand, right hand, and face are evaluated in a unified expressive space.

Let $\mathcal{D}_{\mathrm{SMPLX}}(\cdot)$ denote the expressive SMPL-X decoder that maps the denoised expressive vector to 3D hand joints and facial landmarks:

$$\left(\hat{\mathbf{J}}^{lh}, \hat{\mathbf{J}}^{rh}, \hat{\mathbf{L}}^{face}\right) = \mathcal{D}_{\mathrm{SMPLX}}(\hat{\mathbf{x}}_0). \tag{13}$$

Here, $\hat{\mathbf{J}}^{lh}$ and $\hat{\mathbf{J}}^{rh}$ denote the decoded left- and right-hand joints, while $\hat{\mathbf{L}}^{face}$ denotes the decoded facial landmarks. The geometry-aware loss is defined as:

$$\mathcal{L}_{\mathrm{geo}} = \omega_{lh} \left\| \hat{\mathbf{J}}^{lh} - \mathbf{J}^{lh} \right\|_2^2 + \omega_{rh} \left\| \hat{\mathbf{J}}^{rh} - \mathbf{J}^{rh} \right\|_2^2 + \omega_{face} \left\| \hat{\mathbf{L}}^{face} - \mathbf{L}^{face} \right\|_2^2. \tag{14}$$

The indicators $\omega_{lh}$, $\omega_{rh}$, and $\omega_{face}$ specify whether valid supervision is available for each expressive region. This enables training with heterogeneous samples, including hand-only, face-only, and joint hand–face annotations. The geometry-aware loss directly regularizes the denoised expressive parameters in 3D geometric space, improving finger articulation, jaw motion, and facial landmark consistency.

The final diffusion training objective combines the part-balanced score-matching loss with geometry-aware supervision and optional anatomical regularization:

$$\mathcal{L}_{\mathrm{diff}} = \mathcal{L}_{\mathrm{score}} + \lambda_{\mathrm{geo}} \mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{anat}} \mathcal{L}_{\mathrm{anat}}, \tag{15}$$

where $\lambda_{\mathrm{geo}}$ and $\lambda_{\mathrm{anat}}$ control the contributions of the geometry-aware and anatomical regularization terms, respectively. The anatomical loss is defined as

$$\mathcal{L}_{\mathrm{anat}} = \sum_j \left[ \max\left(0,\, \boldsymbol{\theta}_j - \boldsymbol{\theta}_j^{\max}\right)^2 + \max\left(0,\, \boldsymbol{\theta}_j^{\min} - \boldsymbol{\theta}_j\right)^2 \right] \quad + \left\| \mathbf{b}(\hat{\mathbf{x}}_0) - \mathbf{b}_0 \right\|_2^2 + \left\| \mathbf{x}^{expr} \right\|_2^2, \tag{16}$$

which combines per-joint rotation-limit penalties with limits $[\boldsymbol{\theta}_j^{\min}, \boldsymbol{\theta}_j^{\max}]$, a bone-length consistency term against the template lengths $\mathbf{b}_0$, and an expression-magnitude regularizer. We set $\lambda_{\mathrm{geo}} = 1.0$ and $\lambda_{\mathrm{anat}} = 10^{-3}$ for all experiments. The anatomical loss $\mathcal{L}_{\mathrm{anat}}$ optionally penalizes implausible hand poses, inconsistent bone lengths, and excessive facial expression magnitudes.

### 3.6 Expressive Training and Inference

ExDiff is trained in two stages. ExDiff-B (base) learns independent part priors, a shared MANO Romero et al. (2017) hand prior for both hands and a FLAME-based Li et al. (2017) face prior for jaw pose and expression, each with its own noise schedule (Eq. 4) and no cross-part interaction. ExDiff-F (fused) freezes these priors and trains the zero-initialized fusion head of Eq. 10 on paired hand–face data, capturing dependencies such as mirrored finger configurations and co-occurring jaw–finger articulation. At inference, the learned score $s_\phi(\mathbf{x}_t, t, \mathbf{m}, \mathbf{c})$ serves as a plug-in prior for monocular fitting, where $\mathbf{c}$ weights image evidence against the prior, and for missing-part completion, where parts with $m_p = 0$ are sampled while observed parts are held fixed. Paired data and both procedures are detailed in the Appendix A.1 and A.2.

## 4 Experiments

**Datasets and Implementation Details.** ExDiff is trained on combined hand and face datasets. For hands, we use FreiHAND Zimmermann et al. (2019), DexYCB Chao et al. (2021), HO3D Hampali et al. (2020), H2O Kwon et al. (2021), and ReInterHand Moon et al. (2023), while the face uses WCPA Kao et al. (2022) and the MICA training compilation Zielonka et al. (2022); the paired hand–face data used to train the fusion head of ExDiff-F. Joint rotations are represented in axis-angle format and normalized to zero mean and unit variance following Lu et al. (2025). The model is optimized with Adam (learning rate $2 \times 10^{-4}$, momentum 0.9, stability constant $10^{-8}$). The variance-exploding schedule of Eq. 4 is discretized into $T = 1000$ noise scales with the part-specific ranges $(\sigma_{\min}^h, \sigma_{\max}^h) = (0.01, 50)$ for both hands and $(\sigma_{\min}^{face}, \sigma_{\max}^{face}) = (0.005, 20)$ for the face (Sec. 3.3); $\lambda_p = 1$ for all parts. During sampling, we perform one denoising step per noise scale using stochastic reverse diffusion with a signal-to-noise ratio of 0.16 and final noise removal;

probability-flow sampling is disabled. For evaluation, the batch size is set to 50, and the generated expressive parameters are denormalized and decoded by SMPL-X Pavlakos et al. (2019), keeping only the hand–face vertices as in Cai et al. (2023). Evaluation metrics and computational cost are given in Appendix A.5 and A.6.

### 4.1 Main Results

**Hand–Face Generation and Recovery.** We evaluate ExDiff for both expressive hand–face generation and mesh recovery. For generation, Table 1 shows that all ExDiff variants outperform VPoser, NRDF, and DPoser in overall quality. Although NRDF obtains high diversity and recall, its high FID, low precision, and large dNN indicate poor alignment with the target distribution. In contrast, ExDiff improves realism, coverage, and diversity, with ExDiff-F achieving the best results with the lowest FID 0.04, highest precision 0.83, second highest recall 0.88, lowest dNN 0.96, and second high APD of 6.52. For recovery, we evaluate on ARCTIC Fan et al. (2023) using only hand and face vertices, excluding torso, leg, and arm regions for a fair comparison.

| Methods | APD↑ | FID↓ | Prec.↑ | Rec.↑ | dNN↓ |
|---|---|---|---|---|---|
| VPoser Pavlakos et al. (2019) | 4.82 | 0.71 | 0.27 | 0.39 | 1.86 |
| NRDF He et al. (2024) | 6.74 | 0.78 | 0.05 | 0.94 | 3.08 |
| DPoser Lu et al. (2025) | 5.36 | 0.10 | 0.70 | 0.78 | 1.31 |
| **ExDiff-B** | **5.91** | **0.08** | **0.75** | **0.82** | **1.18** |
| **ExDiff-F** | **6.52** | **0.04** | **0.83** | **0.88** | **0.96** |

Table 1: Quantitative evaluation of joint hand gesture and facial expression generation after excluding torso and leg components.

| Methods | PA-MPVPE ↓ | | |
|---|---|---|---|
| | All (H/F) | Hands | Face |
| GMM Bogo et al. (2016) | 18.20 | 17.93 | 10.92 |
| VPoser-X Pavlakos et al. (2019) | 17.94 | 17.44 | 10.99 |
| DPoser-X Lu et al. (2025) | 15.73 | 15.60 | 9.75 |
| **ExDiff-B** | **15.12** | **14.83** | **9.54** |
| **ExDiff-F** | **13.53** | **12.96** | **8.84** |
| SMPLer-X Cai et al. (2023) | 8.52 | 11.21 | 2.95 |
| + GMM Bogo et al. (2016) | 7.99 | 9.99 | 2.78 |
| + VPoser-X Pavlakos et al. (2019) | 7.70 | 9.50 | 2.83 |
| + DPoser-X Lu et al. (2025) | 6.53 | 7.33 | 2.73 |
| **+ ExDiff-B** | **6.15** | **7.18** | **2.06** |
| **+ ExDiff-F** | **5.12** | **6.01** | **1.47** |

Table 2: Mesh recovery results on ARCTIC Fan et al. (2023). PA-MPVPE is reported for the full expressive mesh (Hand/Face), hands-only, and face-only.

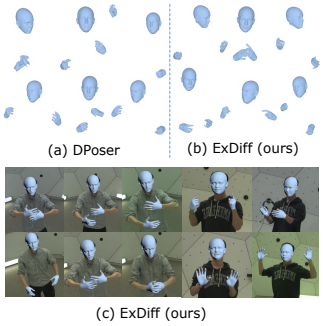


Figure 3: Comparison of hand gestures and facial expressions (a, b) and our (c) plausible gestures and expressions.

As shown in Table 2, ExDiff-B and ExDiff-F consistently outperform GMM, VPoser-X, and DPoser-X when fitting from scratch, and further improve hand–face reconstruction when combined with SMPLer-X Cai et al. (2023) initialization. Qualitative results in Fig. 3 demonstrate that ExDiff generates diverse and plausible hand–face configurations and recovers accurate hand and face geometry from monocular inputs under self-occlusion and motion blur.

**Hand Gesture Generation and Recovery.** We evaluate ExDiff for both hand gesture generation and monocular hand mesh recovery. For generation, Table 3 shows that ExDiff achieves a better balance between diversity, realism, and distribution coverage than VPoser, NRDF, and DPoser. Among the variants, ExDiff-F obtains the best overall results, with the highest APD of 2.90, precision of 0.85, recall of 0.89, and the lowest dNN of 1.05, while maintaining a low FID of 0.02. For recovery, Table 4 shows that all ExDiff variants outperform recent hand mesh recovery methods on FreiHAND Zimmermann et al. (2019). In particular, ExDiff-F achieves the lowest JPE of 4.6 mm and VPE of 5.0 mm, together with the highest F@5 of 0.848 and F@15 of 0.997. The qualitative results in Fig. 4 further confirm that ExDiff produces more natural hand configurations than previous generative models and improves hand mesh recovery on FreiHAND.

| Methods | APD ↑ | FID ↓ | Prec. ↑ | Rec. ↑ | dNN ↓ |
|---|---|---|---|---|---|
| VPoser Pavlakos et al. (2019) | 1.99 | 0.21 | 0.68 | 0.65 | 1.85 |
| NRDF He et al. (2024) | 1.76 | 5.20 | 0.17 | 0.65 | 5.37 |
| DPoser Lu et al. (2025) | 2.36 | **0.01** | 0.82 | 0.87 | 1.45 |
| **ExDiff-B** | **2.61** | **0.05** | **0.83** | **0.87** | **1.24** |
| **ExDiff-F** | **2.90** | **0.02** | **0.85** | **0.89** | **1.05** |

Table 3: Quantitative evaluation of hand gesture generation.

| Methods | JPE ↓ | VPE ↓ | F@5 ↑ | F@15 ↑ |
|---|---|---|---|---|
| Mesh Graph Lin et al. (2021b) | 5.9 | 6.0 | 0.764 | 0.986 |
| MobRecon Chen et al. (2022) | 5.7 | 5.8 | 0.784 | 0.986 |
| AMVUR Jiang et al. (2023) | 6.2 | 6.1 | 0.767 | 0.987 |
| HaMeR Pavlakos et al. (2024) | 6.0 | 5.7 | 0.785 | 0.990 |
| Hamba Dong et al. (2024) | 5.8 | 5.5 | 0.798 | 0.991 |
| NRDF He et al. (2024) | 6.1 | 6.2 | 0.726 | 0.983 |
| DPoser-hand Lu et al. (2025) | 5.3 | 5.2 | 0.793 | 0.990 |
| MaskHand Saleem et al. (2025) | 5.5 | 5.4 | 0.801 | 0.991 |
| HandOS Chen et al. (2025) | 5.0 | 5.3 | 0.812 | 0.991 |
| **ExDiff-B** | **4.8** | **5.2** | **0.827** | **0.994** |
| **ExDiff-F** | **4.6** | **5.0** | **0.848** | **0.997** |

Table 4: Comparison with the state-of-the-art on the FreiHAND dataset Zimmermann et al. (2019). Errors are measured in millimeters.

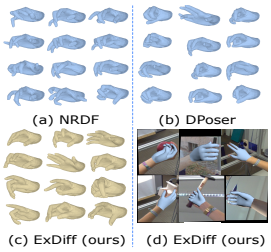


Figure 4: Qualitative hand gesture generation with prior models (a,b) and ours (c,d).

| | Methods | FID ↓ | Prec. ↑ | Rec. ↑ | dNN ↓ |
|---|---|---|---|---|---|
| Shape | VPoser Pavlakos et al. (2019) | 31.91 | 0.984 | 0.105 | 6.52 |
| | DPoser Lu et al. (2025) | 5.331 | 0.689 | 0.396 | 8.29 |
| | **ExDiff-B** | **4.980** | **0.961** | **0.418** | **7.94** |
| | **ExDiff-F** | **4.110** | **0.990** | **0.656** | **5.39** |
| Exp. | VPoser Pavlakos et al. (2019) | 0.888 | 0.993 | 0.019 | 0.79 |
| | DPoser Lu et al. (2025) | 0.156 | 0.818 | 0.697 | 1.01 |
| | **ExDiff-B** | **0.149** | **0.846** | **0.713** | **1.07** |
| | **ExDiff-F** | **0.124** | **0.882** | **0.751** | **0.70** |

Table 5: Quantitative comparison of face generation across shape and expression (exp.).

**Face Expression and Reconstruction.** We evaluate face modeling on shape and expression generation. In Table 5, all ExDiff variants surpass VPoser and DPoser in distribution quality and coverage for both expression and shape, with ExDiff-F giving the best balance of realism, fidelity, and diversity. For the shape, the face branch additionally models the FLAME shape coefficients $\beta$, trained on the MICA compilation Zielonka et al. (2022) with WCPA Kao et al. (2022) expressions; since ARCTIC and BEAT2 provide only 10 SMPL-X shape parameters, ExDiff-F reuses the expression-trained fusion head, so shape gains stem from the face branch rather than fusion.

For reconstruction, we report MPVPE on the NOW benchmark Sanyal et al. (2019) and its side-view subset. Table 6 shows that all ExDiff variants outperform VPoser and DPoser when fitted from scratch, and that MICA Zielonka et al. (2022) initialization further reduces error, with ExDiff-F reaching 6.7/8.7/7.7 mm (median/mean/std) overall and 7.6/9.8/8.0 mm on side views. Fig. 5 confirms diverse generations and realistic reconstructions under occlusion and viewpoint changes.

| Methods | all ↓ | | | side-view ↓ | | |
|---|---|---|---|---|---|---|
| | med. | mea. | std. | med. | mea. | std. |
| VPoser Pavlakos et al. (2019) | 12.2 | 15.1 | 12.7 | 12.2 | 15.4 | 12.9 |
| DPoser Lu et al. (2025) | 11.6 | 14.6 | 12.2 | 11.8 | 14.7 | 12.3 |
| **ExDiff-B** | **11.2** | **14.0** | **11.8** | **11.4** | **14.1** | **11.9** |
| **ExDiff-F** | **10.1** | **12.9** | **10.6** | **10.4** | **13.1** | **12.0** |
| MICA Zielonka et al. (2022) | 9.0 | 11.1 | 9.2 | 9.2 | 11.7 | 10.0 |
| +VPoser Pavlakos et al. (2019) | 9.9 | 12.3 | 10.4 | 10.0 | 12.5 | 10.6 |
| +DPoser Lu et al. (2025) | 8.7 | 10.7 | 9.0 | 9.1 | 11.4 | 9.7 |
| **+ExDiff-B** | **8.2** | **10.1** | **8.8** | **8.7** | **11.0** | **9.3** |
| **+ExDiff-F** | **6.7** | **8.7** | **7.7** | **7.6** | **9.8** | **8.0** |

Table 6: Face reconstruction performance on the NOW benchmark Sanyal et al. (2019). Results are reported as the median, mean, and standard deviation of MPVPE.

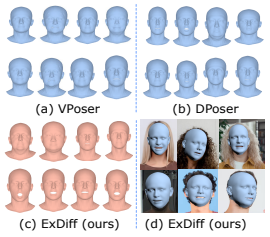


Figure 5: Qualitative facial-expression generation. Previous models (a,b) and ours (c,d).

## 4.2 Ablation Study

**Random vs. Part-Aware Noise Schedule.** Table 7 compares random and part-aware noise schedules on ARCTIC Fan et al. (2023). Across all ExDiff variants, the part-aware schedule consistently reduces PA-MPVPE for both hands and face, demonstrating the benefit of assigning noise according to the characteristics of each expressive region. The best performance is achieved by ExDiff-F with the part-aware schedule, obtaining 5.12 overall PA-MPVPE, 6.01 for hands, and 1.47 for face. These results show

| Noise Schedule | Variant | PA-MPVPE ↓ | | |
|---|---|---|---|---|
| | | All (H/F) | Hands | Face |
| **Random** | **ExDiff-B** | **8.72** | **9.57** | **4.85** |
| | **ExDiff-F** | **7.31** | **8.18** | **3.56** |
| **Part-aware** | **ExDiff-B** | **6.15** | **7.18** | **2.06** |
| | **ExDiff-F** | **5.12** | **6.01** | **1.47** |

Table 7: Ablation study of random and part-aware noise schedules on ARCTIC dataset.

that part-aware noise scheduling provides more effective denoising supervision than random noise injection, especially for fine-grained facial and hand reconstruction.

**Cross-part and Geometry-aware Supervision.** Fig. 6 ablates the two components on BEAT2 Liu et al. (2024), where evaluation is restricted to the hand and face regions; each of the four configurations is a separate run, and the base priors without fusion or $\mathcal{L}_{\text{geo}}$ serve as the reference (14.53/15.43/11.47 mm for All/Hands/Face). Cross-part fusion, which models dependencies among the left hand, right hand, and face, reduces these errors by 21.5%, 22.2%, and 20.9% on its own, and by 29.4%, 30.7%, and 25.0% once geometry-aware supervision is present, i.e., from EXDIFF-B (9.24/10.15/6.03) to EXDIFF-F (6.52/7.03/4.52). Geometry-aware supervision, which regularizes the denoised parameters in 3D hand-joint and facial-landmark space, is the larger single contributor: it lowers the errors by 36.4%, 34.2%, and 47.4% without fusion and by 42.8%, 41.5%, and 50.2% with it. The two components are complementary, since each yields a larger gain when the other is enabled, and together they reduce the overall, hand, and face errors by 55.1%, 54.4%, and 60.6% relative to the base priors. Cross-part reasoning thus improves coordinated hand–face modeling, while geometry-aware supervision provides most of the spatial accuracy for fine-grained expressive reconstruction.

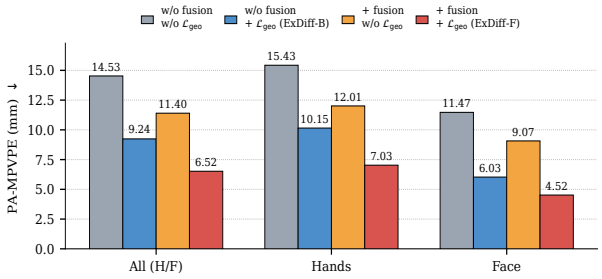


Figure 6: Ablation study of cross-part residual and geometry-aware denoising supervision for expressive mesh recovery on BEAT2 Liu et al. (2024).

**Expressive Training Strategy.** We ablate the expressive training strategy in Table 8. The hand-only and face-only priors are already strong on their own parts (precision 0.83 and 0.846), yet EXDIFF(face/hand)-B, which concatenates them without cross-part interaction, drops to a joint precision of 0.75, since independently sampled parts rarely co-occur. Training the fusion head on paired data, EXDIFF(face/hand)-F, restores part-level fidelity while modeling coordination, achieving the best joint FID of 0.04, precision of 0.83, and recall of 0.88. The proposed strategy therefore improves joint distribution modeling and yields more coherent hand–face generation.

| Method | FID ↓ | Precision ↑ | Recall ↑ |
|---|---|---|---|
| **EXDIFF(face/hand)-B** | **0.08** | **0.75** | **0.82** |
| **EXDIFF(face/hand)-F** | **0.04** | **0.83** | **0.88** |
| **EXDIFF-hand** | **0.05** | **0.83** | **0.87** |
| **EXDIFF-face** | **0.149** | **0.846** | **0.713** |

Table 8: Ablation of the expressive training strategy. Joint hand–face EXDIFF variants are compared with hand-only and face-only generation.

## 5 CONCLUSION AND LIMITATIONS

We presented EXDIFF, an expressive diffusion prior that models the three most communicative regions of the body, specifically the left hand, right hand, and face, instead of learning a full-body prior. EXDIFF combines a part-balanced score-matching objective with part-aware noise schedules, a part-aware denoising network with zero-initialized cross-part residual fusion, and geometry-aware denoising supervision through SMPL-X decoding. Trained on heterogeneous hand-only, face-only, and joint hand–face data with part-availability masks, a single model supports expressive generation, missing-part completion, and monocular recovery.

**Limitations.** Because hands and face are modeled in wrist- and head-local frames on a fixed body, EXDIFF captures dependencies between local articulations but not hand-to-head placement (e.g., touching the face); adding a wrist-to-head transform is the most direct extension, and the continuous 6D rotation representation Zhou et al. (2019) is a natural drop-in for the axis-angle parameterization. The prior is also static and per-frame, and is trained mostly on manipulation data without speech or body context, so it serves as a parametric pose prior rather than a model of communicative behavior.

## REFERENCES


Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6121–6131, 2020.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pp. 561–578. Springer, 2016.

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*, pp. 10843–10852, 2019.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12417–12426, 2021.

Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9087–9098, 2023.

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*, pp. 9044–9053, 2021.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13274–13283, 2021a.

Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20544–20554, 2022.

Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. Handos: 3d hand reconstruction in one stage. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17304–17314, 2025.

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*, pp. 10451–10460, 2021b.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20311–20322, 2022.

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–13, 2023.

Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12943–12954, 2023.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10833–10842, 2019.

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*, pp. 3196–3206, 2020.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11807–11816, 2019.

Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1661–1671, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 118–134, 2018.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11107–11116, 2021.

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*, pp. 758–767, 2023.

Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 2017.

Yueying Kao, Bowen Pan, Miao Xu, Jiangjing Lyu, Xiangyu Zhu, Yuanzhang Chang, Xiaobo Li, Zhen Lei, and Zixiong Qin. Single-image 3d face reconstruction under perspective projection. *arXiv preprint arXiv:2205.04126*, 3(8), 2022.

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*, pp. 12880–12889, 2023.

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*, pp. 4990–5000, 2020.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10138–10148, 2021.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 394–403, 2023.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Kevin Lin, Lijuan Wang, Zicheng Liu, 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*, pp. 1954–1963, 2021a.

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

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1144–1154, 2024.

Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14687–14697, 2021.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, et al. Dposer-x: Diffusion model as robust 3d whole-body human pose prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9988–9997, 2025.

Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pp. 752–768. Springer, 2020.

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 *European Conference on Computer Vision*, pp. 548–564. Springer, 2020.

Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *Advances in Neural Information Processing Systems*, 36:17689–17701, 2023.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

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*, pp. 9826–9836, 2024.

George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2490–2501, 2024.

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), November 2017.

Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. Maskhand: Generative masked modeling for robust hand mesh reconstruction in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8372–8383, 2025.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7763–7772, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pp. 160–177. Springer, 2022.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Linlin Yang, Shicheng Chen, and Angela Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11364–11373, 2021a.

Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint arXiv:2008.05079*, 2020.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11097–11106, 2021b.

Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11354–11363, 2021a.

Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. Accurate 3d face reconstruction with facial component tokens. in 2023 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pp. 8999–9008, 2023.

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*, pp. 2354–2364, 2019.

Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11281–11292, 2021b.

Zimeng Zhao, Xi Zhao, and Yangang Wang. Travelnet: Self-supervised physically plausible hand motion learning from monocular color images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11666–11676, 2021.

Ce Zheng, Xianpeng Liu, Qucheng Peng, Tianfu Wu, Pu Wang, and Chen Chen. Diffmesh: A motion-aware diffusion framework for human mesh recovery from videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4891–4901. IEEE, 2025.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.

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*, pp. 1367–1376, 2024.

Wojciech Zielonka, Timo Bolkart, Justus Thies, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pp. 250–269. Springer, 2022.

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*, pp. 813–822, 2019.

# A APPENDIX

This appendix provides the details needed to reproduce the results in the main text. Appendix A.1 specifies the paired hand–face training data of EXDIFF-F and the ablation variants of Fig. 6. Appendix A.2 gives the monocular fitting and missing-part completion procedures, and Appendix A.3 shows inverse kinematics under partial and noisy 3D observations. Appendix A.4 describes the face–hand submesh used for evaluation and visualization, Appendix A.5 reports computational cost, Appendix A.6 defines the evaluation metrics and generation protocol, Appendix A.7 specifies the face-shape variant of Table 5, and Appendix A.8 provides additional qualitative generations.

## A.1 PAIRED DATA AND ABLATION VARIANTS

**Paired data.** The fusion head of EXDIFF-F is trained on the paired hand–face annotations of the official *train* splits of ARCTIC Fan et al. (2023) and BEAT2 Liu et al. (2024). After 10-frame temporal subsampling this yields ≈160k frames from 38 subjects (8 ARCTIC, 30 BEAT2). The SMPL-X hand poses, jaw pose, and expression coefficients are converted to the wrist- and head-local frames of Sec. 3.1 and normalized with the shared normalizer of Sec. 3.2; the 10 SMPL-X shape parameters provided by these datasets are not used (Appendix A.7). All reconstruction and ablation results on ARCTIC and BEAT2 use the official *test* splits, which share no subjects or sequences with the fusion training data.

**Ablation variants.** The four configurations form a $2\times2$ grid over the two components of EXDIFF-F. "w/o fusion" is EXDIFF-B: the frozen part priors of Eq. 8 with no residual head, so no information is exchanged between parts. "+ fusion" adds the zero-initialized head of Eq. 10, trained on the paired data above. "w/o $\mathcal{L}_{\text{geo}}$" sets $\lambda_{\text{geo}} = 0$ in Eq. 15 in both training stages; "+ $\mathcal{L}_{\text{geo}}$" uses $\lambda_{\text{geo}} = 1$. Hence "w/o fusion + $\mathcal{L}_{\text{geo}}$" is EXDIFF-B and "+ fusion + $\mathcal{L}_{\text{geo}}$" is EXDIFF-F as reported in every other table. Each configuration is a single training run evaluated with the fitting pipeline of Appendix A.2 and SMPLer-X initialization. The random-schedule rows of Table 7 replace the part-specific ranges of Sec. 3.3 with a single range $(\sigma_{\min}, \sigma_{\max}) = (0.01, 50)$ for all parts.

## A.2 MONOCULAR FITTING AND MISSING-PART COMPLETION

**Observations.** Given an image, 2D whole-body keypoints (body, both hands, face) are obtained from ViTPose-WholeBody Xu et al. (2022); each keypoint $\mathbf{j}_k$ carries a detector confidence $c_k \in [0, 1]$. The part confidence $\mathbf{c} = (c_{lh}, c_{rh}, c_{face})$ used for conditioning is the mean keypoint confidence of that part, and the availability mask is $m_p = \mathbf{1}[c_p > \tau]$ with $\tau = 0.3$.

**Camera and body.** Following SMPLify-X Pavlakos et al. (2019), a perspective camera with focal length 5000 px, the global orientation and translation, and the body pose (neutral when fitting from scratch, or taken from SMPLer-X Cai et al. (2023)) are estimated from the body keypoints in a first stage; none of these enter $\mathbf{x}$ (Sec. 3.1).

**Objective.** The expressive vector is then optimized by

$$E(\mathbf{x}) = \sum_k c_k \, \rho\big(\Pi\big(\mathcal{J}_k(\mathbf{x})\big) - \mathbf{j}_k\big) + \lambda_{\text{prior}} \, \mathcal{L}_{\text{prior}}(\mathbf{x}) + \lambda_{\text{anat}} \, \mathcal{L}_{\text{anat}}(\mathbf{x}), \tag{17}$$

where $\Pi$ projects the SMPL-X joints $\mathcal{J}_k$, $\rho$ is the Geman–McClure robustifier, and $\mathcal{L}_{\text{anat}}$ is Eq. 16 with $\lambda_{\text{anat}} = 10^{-3}$. Following DPoser-X Lu et al. (2025), the prior term perturbs the current estimate to a time $t \sim \mathcal{U}(t_{\min}, t_{\max})$ with the part schedules of Eq. 4, denoises it with Eq. 11, and penalizes

$$\mathcal{L}_{\text{prior}}(\mathbf{x}) = \big\|\mathbf{x} - \text{sg}\big[\hat{\mathbf{x}}_0(\mathbf{x}_t, t)\big]\big\|_2^2, \tag{18}$$

with the gradient stopped (sg) through the network.

**Optimization.** We use Adam with learning rate $10^{-2}$ for 300 iterations in 3 stages of 100 iterations, annealing $t_{\max}$ from 0.6 to 0.1 with $t_{\min} = 0.02$, and $\lambda_{\text{prior}} = 10^{-2}$. "From scratch" rows initialize $\mathbf{x}$ at the training-set mean; "+SMPLer-X" and "+MICA" rows initialize from the regressor output and keep the same objective and iteration budget. The identical procedure is used for GMM,

VPoser, and DPoser-X by swapping $\mathcal{L}_{\text{prior}}$, so Tables 2, 4, and 6 compare priors under one fitting pipeline; direct image-to-mesh regressors in Table 4 are reported as published, and the prior rows of that table (NRDF, DPoser-hand, ExDiff) are fitted on top of the HaMeR Pavlakos et al. (2024) initialization.

**Completion.** For missing-part completion the observed parts are fixed and the missing parts are sampled from the prior: at every reverse step the observed part tokens are replaced by their forward-noised ground truth at the current $t$ (as in inpainting), while the missing parts follow the conditional score of Eq. 5 with the corresponding $\mathbf{m}$ and $\mathbf{c} = \mathbf{1}$. The same mechanism is used when a part falls below the confidence threshold $\tau$ during fitting.

### A.3 Inverse Kinematics from Partial 3D Observations

We further evaluate ExDiff as a diffusion-based prior for inverse kinematics from partial and noisy 3D observations. Given sparse 3D evidence, the goal is to recover plausible hand articulation or facial expression parameters that remain consistent with the observed joints or landmarks. For a hand $h \in \{lh, rh\}$ with observed 3D joints $\mathbf{J}^{obs}$ and visibility mask $\mathbf{M}_J$, we optimize

$$\mathbf{x}^{h*} = \arg\min_{\mathbf{x}^h} \left\| \mathbf{M}_J \odot \left( \mathbf{J}(\mathbf{x}^h) - \mathbf{J}^{obs} \right) \right\|_2^2 + \lambda_{\text{prior}} \, \mathcal{L}_{\text{prior}}(\mathbf{x}^h), \tag{19}$$

where $\mathcal{L}_{\text{prior}}$ is the prior term of Eq. 18 applied to the normalized hand pose, with the same optimizer settings as Appendix A.2. The first term fits the visible 3D joints, while the prior regularizes the solution toward realistic articulation. The face case is identical with 3D landmarks $\mathbf{L}^{obs}$ and the face vector $\mathbf{x}^{face}$.

**Hand Inverse Kinematics.** We evaluate on ReInterHand Moon et al. (2023), dropping 30% of the joints uniformly at random and adding Gaussian noise of $\sigma = 10$ mm to the rest. As shown in Fig. 7, ExDiff recovers plausible hand articulations even when the observed joints are sparse or corrupted. The recovered meshes remain anatomically consistent and closely aligned with the available joint evidence.

**Face Inverse Kinematics.** We evaluate on WCPA Kao et al. (2022) under the same corruption. Fig. 8 shows that ExDiff reconstructs realistic and well-aligned facial geometry despite missing or perturbed landmarks, preserving plausible facial structure and expression detail.

### A.4 Face–Hand Submesh Extraction and Rendering

After optimization or generation, the predicted expressive parameters are decoded with SMPL-X:

$$\mathbf{V}, \mathcal{F} = M_{\text{SMPLX}}(\Theta), \tag{20}$$

where $\mathbf{V}$ and $\mathcal{F}$ denote the full SMPL-X vertex set and triangle topology. Since ExDiff models only the expressive regions, we extract a compact face–hand submesh using a predefined vertex index set $\mathcal{I}_{fh}$:

$$\mathbf{V}_{fh} = \mathbf{V}[\mathcal{I}_{fh}], \qquad \mathcal{F}_{fh} = \{ f_j \in \mathcal{F} \mid f_{j,1}, f_{j,2}, f_{j,3} \in \mathcal{I}_{fh} \}. \tag{21}$$

The selected faces are re-indexed to the compact vertex set, which prevents vertex–face mismatch and allows rendering only the hands and face while preserving their positions in SMPL-X space. $\mathcal{I}_{fh}$ is the union of the official SMPL-X hand and FLAME vertex index sets ($|\mathcal{I}_{fh}| = 6{,}579$ vertices: 778 per hand and 5,023 for the face) and defines the "All", "Hands", and "Face" regions over which PA-MPVPE is computed in Tables 2 and 7 and Fig. 6.

For qualitative analysis, the submesh is rendered from canonical viewpoints by rotating it about its centroid,

$$\mathbf{V}_{fh}^{view} = R \left( \mathbf{V}_{fh} - \bar{\mathbf{V}}_{fh} \right) + \bar{\mathbf{V}}_{fh}, \tag{22}$$

where $R$ is the view rotation and $\bar{\mathbf{V}}_{fh}$ the centroid. We render front, back, top, bottom, left, and right views.

Fig. 9 shows renderings on CMU Panoptic Joo et al. (2017); the submesh preserves the spatial arrangement between the face and both hands while removing full-body geometry. This arrangement

is inherited from the fitted body scaffold and camera, not from the diffusion prior, which models only wrist- and head-local articulation (Sec. 3.1). Fig. 10 presents single-view results on in-the-wild images, and Fig. 11 hand–face recovery on sign-language and random images.

| ID | Name | Type | Parameters | GFLOPs | Runtime (ms) |
|---|---|---|---|---|---|
| 0a | part priors (shared hand $s_h$, face $s_f$) | TimeFC | 5.5M | ≈0.05 | ≈20.0 |
| 0b | cross-part fusion blocks + residual head ($L = 4$, width 256, 4 heads) | attention | 2.1M | ≈0.02 | ≈7.5 |
| 0 | ExDiff hand–face denoiser (0a+0b) | | 7.6M | ≈0.07 | ≈27.5 |
| 1 | hand-model-visual | MANO | 3.2K | ≈0.0001 | ≈0.6 |
| 2 | hand-model-eval | MANO | 16.0K | ≈0.0002 | ≈0.6 |
| 3 | face-model-visual | FLAME | 8.1K | ≈0.0003 | ≈0.8 |
| 4 | face-model-eval | FLAME | 32.0K | ≈0.0005 | ≈1.0 |
| Trainable parameters | | | 7.6M | | |
| Non-trainable parameters | | | 59.3K | | |
| Total parameters | | | 7.66M | ≈0.07 | ≈30.5 |
| Estimated model size (MB) | | | 30.64 | | |

Table 9: Model architecture, parameter statistics, computational complexity, and runtime for ExDiff hand–face generation. Rows 0a–0b split the trainable denoiser into the TimeFC part priors (Eq. 8) and the $L$ attention-based cross-part fusion blocks with the zero-initialized residual head (Eqs. 9–10); runtime is per denoising step at batch size 50 on a single NVIDIA A100 GPU (a full $T = 1000$-step sample therefore takes ≈30 s per batch). MANO and FLAME are non-trainable mesh decoders used for visualization and evaluation.

### A.5 Computational Cost Analysis

Table 9 reports the computational cost of ExDiff for expressive hand–face generation. Unlike hand-only generation, ExDiff predicts a unified expressive representation that includes both hand gestures and facial expression parameters, so the profile includes the trainable hand–face denoiser together with non-trainable mesh decoders for both hands and face.

The main trainable component is the hand–face denoiser, consisting of the TimeFC part priors for the shared hand branch and the face branch (row 0a) and the $L$ cross-part attention blocks with the zero-initialized residual head (row 0b). It contains approximately 7.6M trainable parameters and requires about 0.07 GFLOPs per sample. Compared with the hand-only setting, the parameter count and runtime increase slightly because the network predicts a larger representation that includes hand pose, jaw pose, and facial expression coefficients.

The MANO and FLAME modules are used only as non-trainable decoders for visualization and evaluation: MANO decodes the generated hand parameters into hand meshes, and FLAME decodes the facial parameters into face geometry and landmarks. They add only 59.3K non-trainable parameters, negligible GFLOPs, and roughly 3.0 ms per sample.

Overall, ExDiff contains approximately 7.66M parameters, of which 7.6M are trainable, an estimated model size of 30.64 MB, and a runtime of about 30.5 ms per denoising step. The proposed model therefore remains computationally lightweight while supporting joint hand-gesture and facial-expression generation; the additional cost of the face branch and fusion head is modest.

### A.6 Evaluation Metrics and Protocol

We evaluate ExDiff with task-specific metrics for expressive hand–face generation and reconstruction. All reconstruction errors are reported in millimeters.

**Expressive Hand–Face Generation.** For generation quality, we report Fréchet Inception Distance (FID) Heusel et al. (2017), Precision, and Recall Kynkäänniemi et al. (2019). FID measures the distribution gap between generated and real expressive samples, where lower is better. Precision evaluates the fidelity of generated samples, while Recall measures the coverage and diversity of the generated distribution. We also report Average Pairwise Distance (APD) and dNN He et al. (2024): APD measures diversity among generated samples, while dNN measures the distance between each

generated sample and its nearest neighbor in the training set, indicating whether generated results remain close to the real expressive manifold.

**Generation protocol.** All generation metrics are computed in the normalized expressive parameter space of Sec. 3.2, following DPoser-X Lu et al. (2025). Tables 1, 3, and 5 use 5,000 generated samples against 5,000 real samples drawn from the ARCTIC, FreiHAND, and WCPA test splits, respectively; precision and recall use the $k$-NN manifold estimate with $k = 3$, dNN is the mean generated-to-real nearest-neighbor distance, and APD the mean pairwise distance among 500 generated samples. Table 8 uses the same protocol: its joint rows are the EXDIFF-B/-F rows of Table 1, and its part-only rows are the EXDIFF-B rows of Tables 3 and 5 (Exp.), evaluated against their own part distribution. Baselines (VPoser Pavlakos et al. (2019), NRDF He et al. (2024), DPoser-X Lu et al. (2025)) are evaluated with the released whole-body models, truncated to their hand and face components, and the same evaluation code is applied to every method. Throughout the paper, "DPoser" in generation tables and "DPoser-X" in recovery tables refer to the same released DPoser-X model.

**Expressive Mesh Recovery.** We report Procrustes-aligned Mean Per-Vertex Position Error (PA-MPVPE), the vertex-level error after rigid alignment, separately for All, Hands, and Face, where All denotes the combined hand–face region defined by $\mathcal{I}_{fh}$ in Appendix A.4 rather than the full body.

**Hand Gesture Reconstruction.** For FreiHAND we report Joint Position Error (JPE), Vertex Position Error (VPE), and F-scores at 5 and 15 mm, all after Procrustes alignment as in the official benchmark.

**Face Expression Reconstruction.** For NOW we report the median, mean, and standard deviation of the scan-to-mesh distance under the official protocol, for all views and for the side-view subset.

### A.7 FACE-SHAPE GENERATION VARIANT

For the shape rows of Table 5, the face branch operates on $[\mathbf{x}^{jaw}; \mathbf{x}^{expr}; \boldsymbol{\beta}]$ with FLAME Li et al. (2017) shape coefficients $\boldsymbol{\beta} \in \mathbb{R}^{N_\beta}$, $N_\beta = 300$, normalized with their own per-dimension statistics. $\boldsymbol{\beta}$ shares the face schedule of Eq. 4 and enters the part-balanced loss of Eq. 6 as an additional term with $D_\beta = N_\beta$; no geometry supervision is applied to shape. The shape model is trained on the FLAME shape coefficients of the MICA compilation Zielonka et al. (2022) (LYHM, FRGC, D3DFACS, Stirling, FaceWarehouse, and Florence), with WCPA Kao et al. (2022) expressions. Because ARCTIC and BEAT2 provide only 10 SMPL-X shape parameters, EXDIFF-F reuses the expression-trained fusion head unchanged in the shape rows, so shape gains stem from the face branch rather than fusion. Shape is excluded from the joint hand–face representation of Eq. 1 in all other experiments.

### A.8 ADDITIONAL QUALITATIVE GENERATIONS

Figure 12 presents joint hand–face generations, showing that the model synthesizes jointly sampled combinations of bilateral hand gestures and facial expressions within a unified expressive space; the placement of the hands relative to the head is fixed by the neutral scaffold rather than generated.

Fig. 13 shows additional left- and right-hand samples with diverse finger articulations across both hands, indicating that the prior captures a broad range of plausible gestures rather than collapsing to a few common poses. Fig. 14 highlights generated facial variations involving both eye and mouth regions, showing that EXDIFF models coupled facial variations beyond isolated local movements. Fig. 15 focuses on mouth-only expression generation, and Fig. 16 presents additional plausible face generations with diverse expression patterns and natural structure.

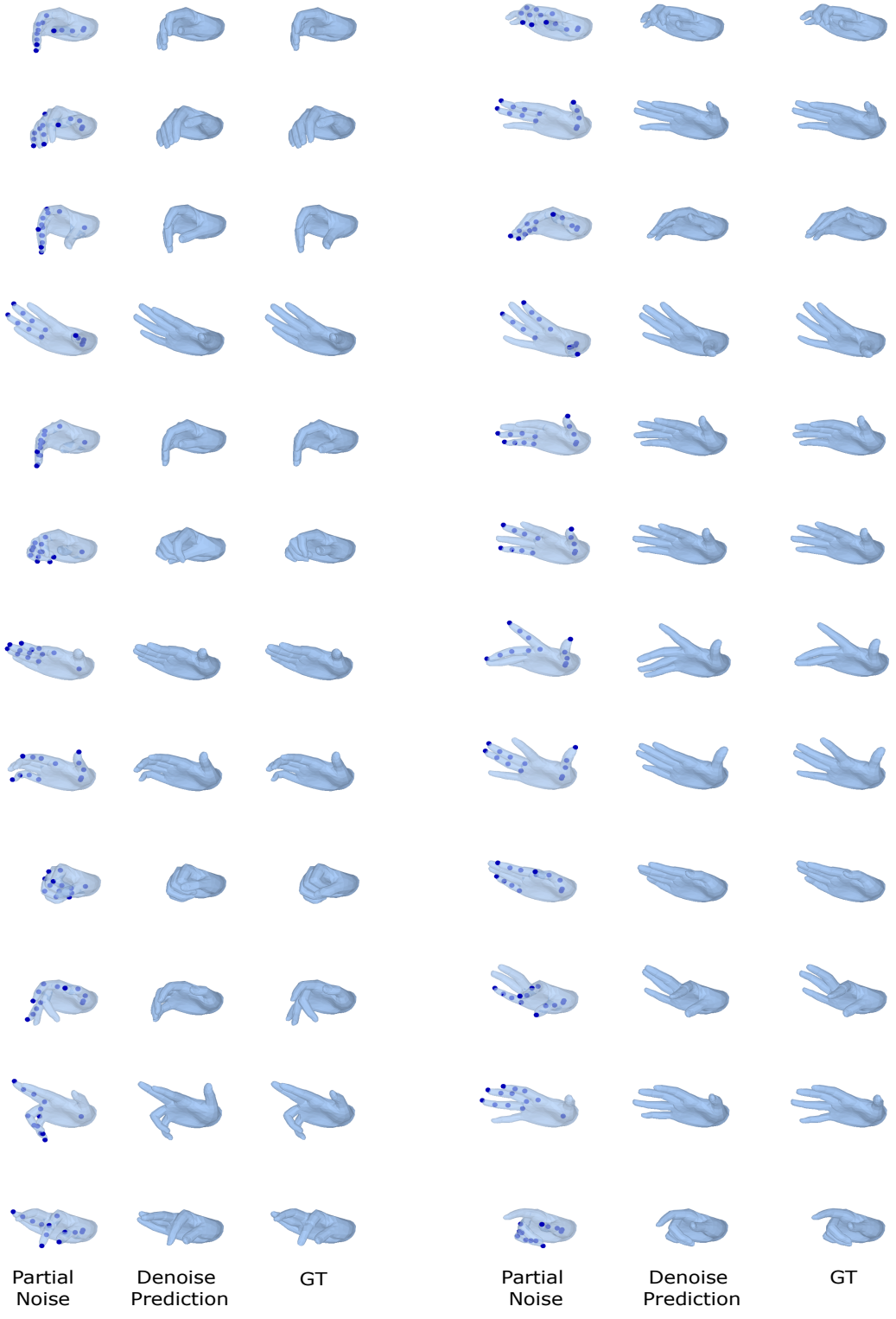


Figure 7: Qualitative results of hand inverse kinematics on ReInterHand Moon et al. (2023) under partial and noisy 3D joint observations.

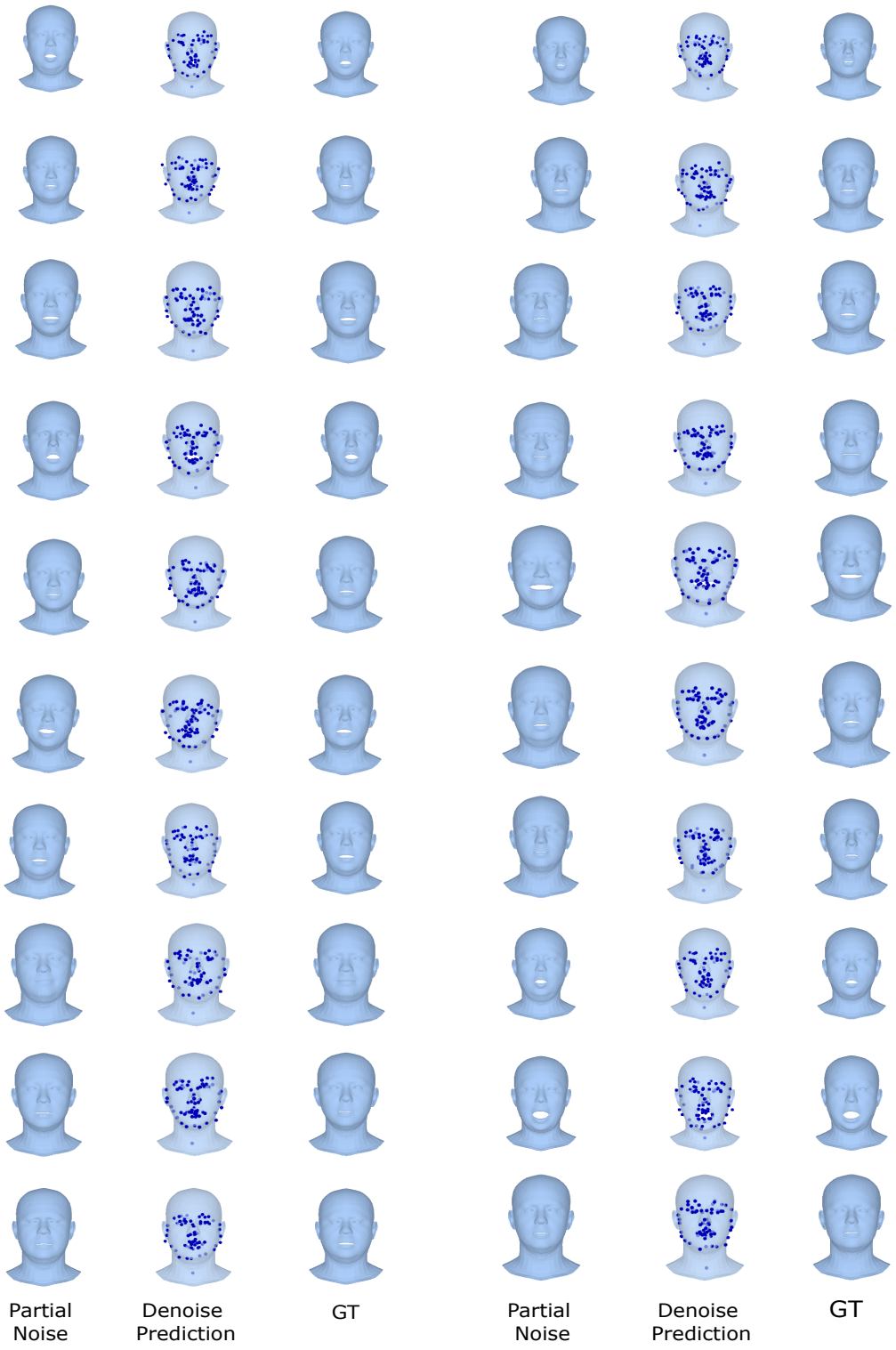


Figure 8: Qualitative results of face inverse kinematics on WCPA Kao et al. (2022) under noisy and occluded landmark observations.

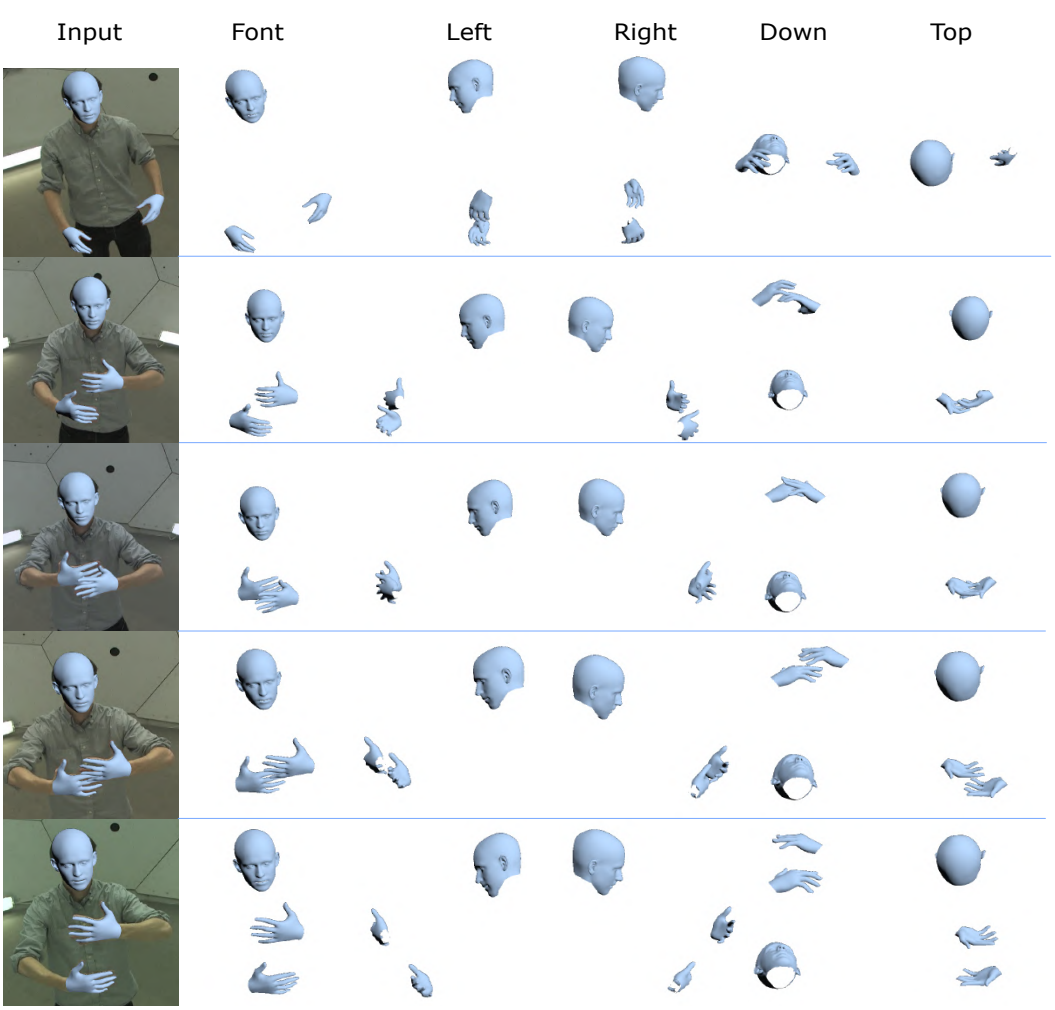


Figure 9: Face–hand mesh rendering on samples from CMU Panoptic Joo et al. (2017). The extracted submesh preserves the spatial relationship between the face and both hands as given by the fitted scaffold and camera, while removing full-body geometry.



Figure 10: Single-view qualitative results of hand-gesture and facial-expression rendering on in-the-wild images.

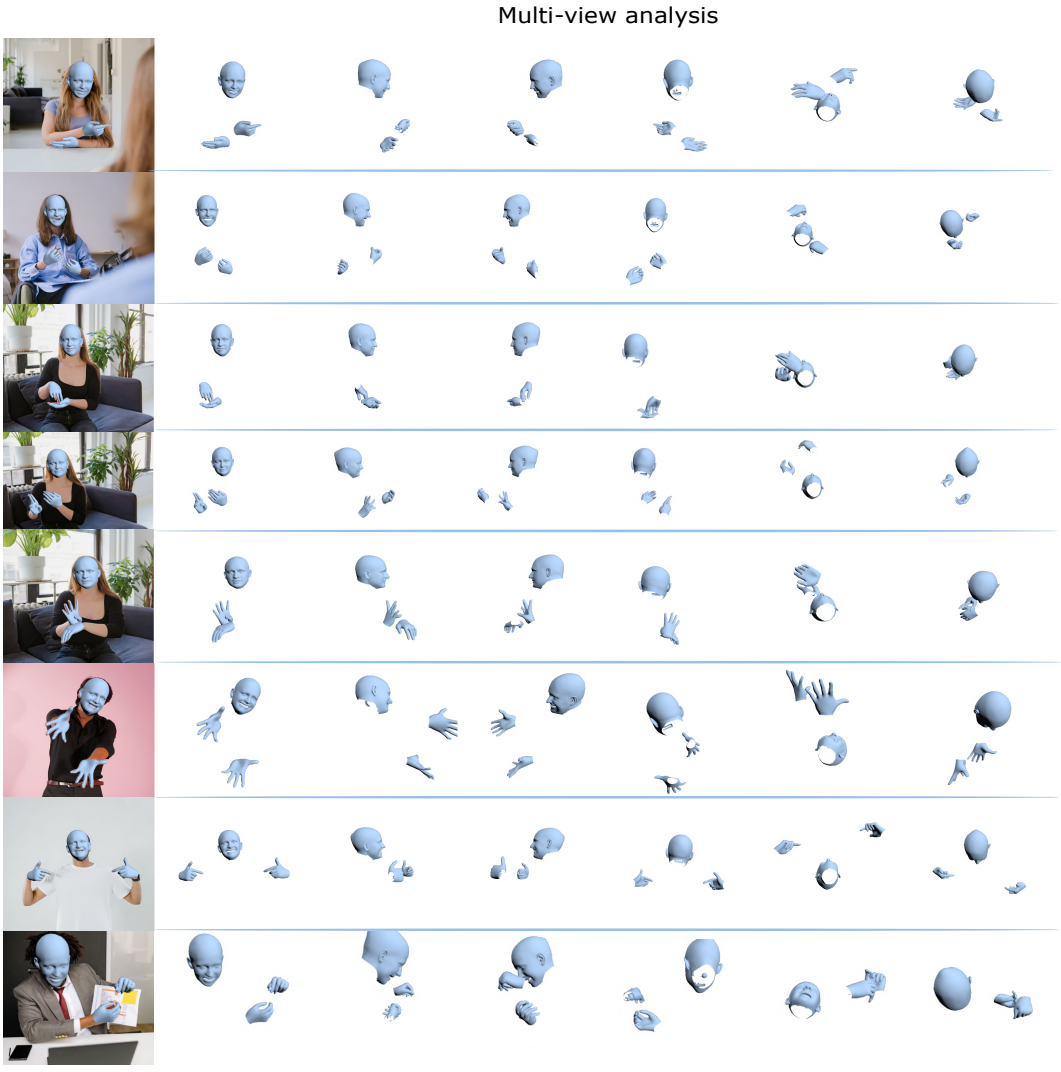


Figure 11: Qualitative hand–face mesh recovery on sign-language and random images, showing joint reconstruction of expressive hand gestures and facial regions. Best viewed with zoom.

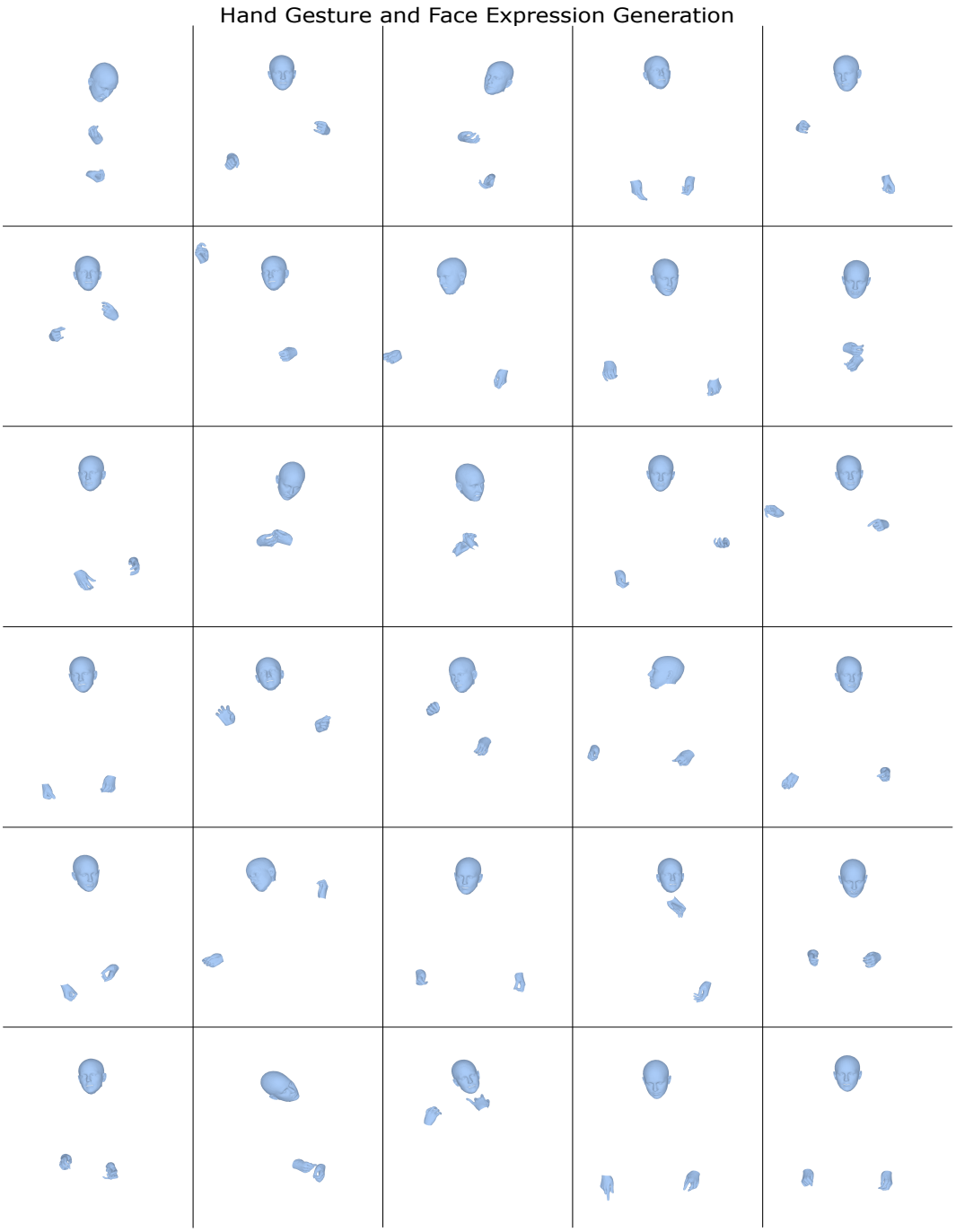


Figure 12: Additional joint hand–face generations. EXDIFF produces jointly sampled bilateral hand gestures and diverse facial expressions in a unified expressive space; hand placement relative to the head follows the neutral scaffold.

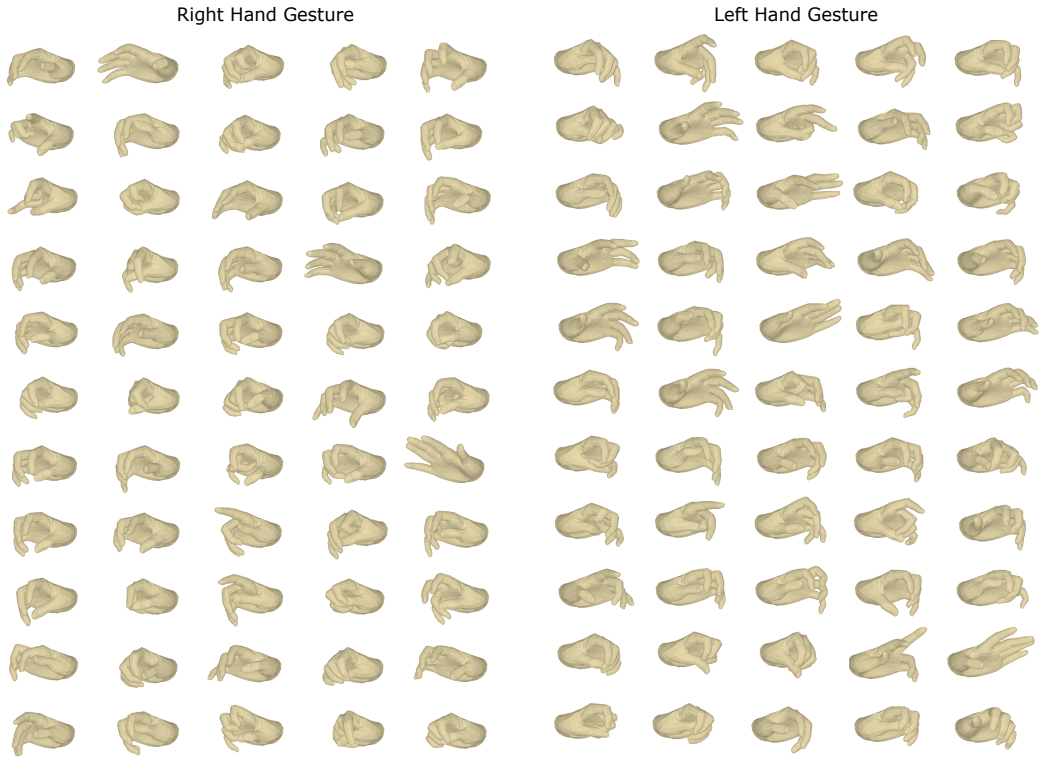


Figure 13: Additional left- and right-hand generations, illustrating diverse and plausible articulations across both hands.

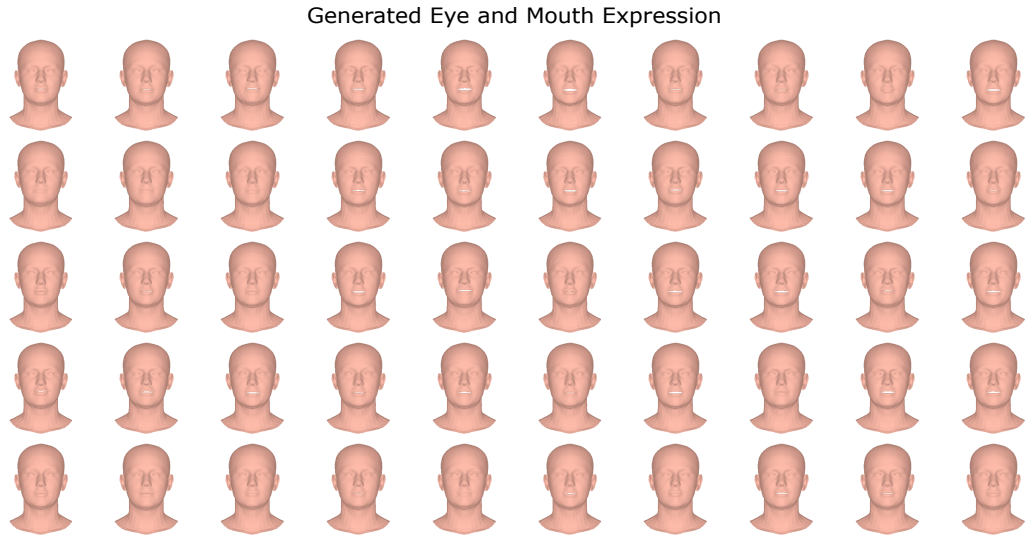


Figure 14: Generated facial variations involving the eye and mouth regions, showing coupled facial variations and diverse expressive combinations.

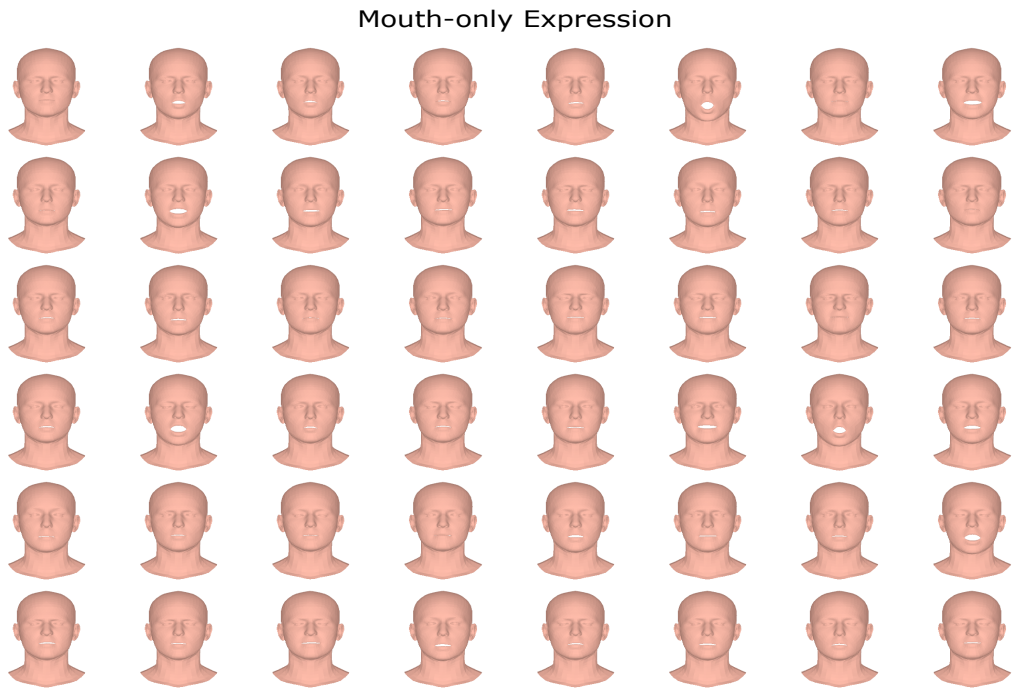


Figure 15: Generated mouth-only expressions with varied, realistic mouth shapes and plausible facial structure.

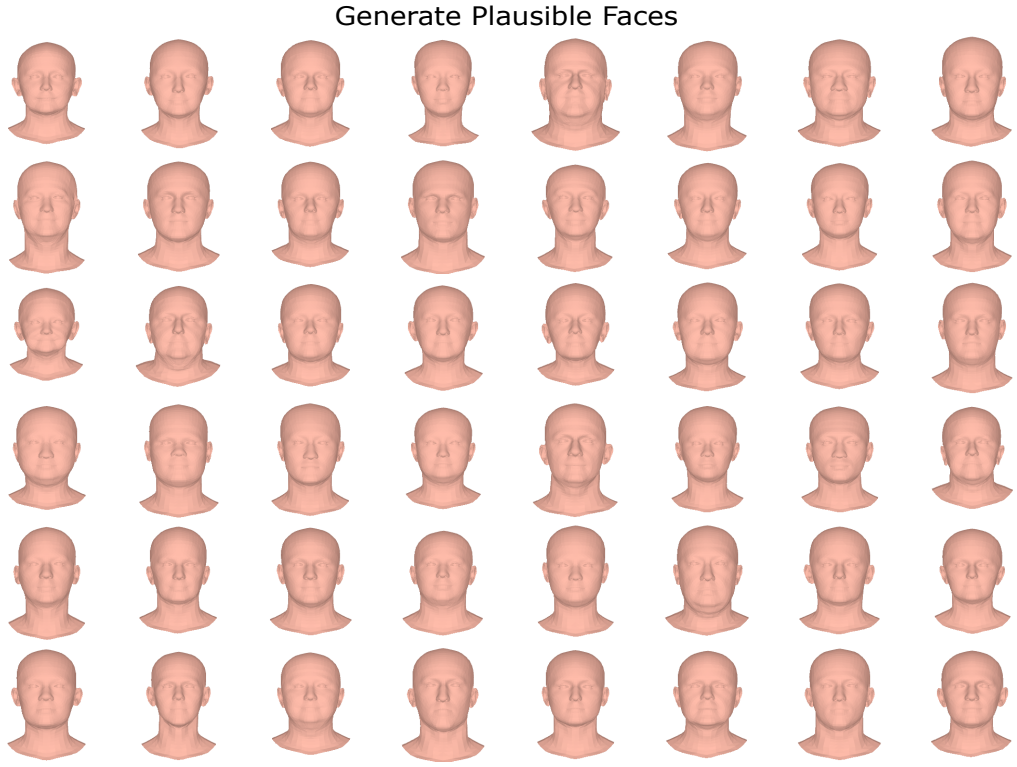


Figure 16: Additional plausible face generations with diverse facial geometry and expression patterns.