GRAPE: GRADUATED ROUTING FOR ARTICULATED PORTRAIT MESH ESTIMATION

Anonymous authors
Paper under double-blind review

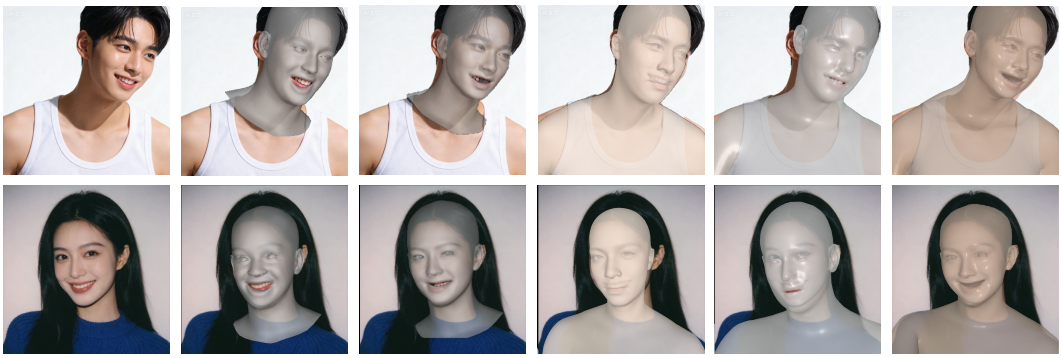


Figure 1: We propose **GRAPE** (Graduated Routing for Articulated Portrait mesh Estimation). The input portrait images are AI generated.

ABSTRACT

Articulated portrait mesh estimation is fundamental to 3D understanding, avatar generation, and immersive interaction. Existing approaches primarily rely on 3D Morphable Models (3DMMs). However, face-centric models suffer from the "floating head" assumption, conflating head pose with global rotation due to the lack of neck kinematics. Conversely, body-centric models lack high-fidelity facial expression capabilities. Furthermore, current methods often let expression blendshapes shortcut jaw opening. These limitations make monocular portrait recovery difficult across representation, supervision, and anatomical parameter estimation. To address these limitations, we introduce GRAPE (Graduated Routing for Articulated Portrait mesh Estimation). We formulate a Portrait Parametric Model (PPM) as a task-specific articulated portrait representation: a torso-rooted kinematic reference for joint head–torso estimation, instantiated with a FLAME head on an SMPL-X torso. We propose a Progressive Anatomical Alignment (PAA) network, which is composed of a pretrained portrait encoder, a Graduated-Mask Router, and coarse-to-fine experts that follow this kinematic order. We then train this network with multi-source supervision that combines sparse anatomical keypoints, feature distillation, foreground mask constraints, and relative geometry constraints. Experiments show that GRAPE improves pose alignment and jaw–skeletal consistency, and reduces jaw–expression shortcutting, while head/face vertex error stays comparable to strong face regressors. The estimated parameters also improve the training/reconstruction behavior of downstream talking-head and avatar models.

1 INTRODUCTION

Articulated portrait mesh estimation is pivotal for 3D scene understanding, immersive telepresence, and the animation of digital avatars. While 3D Morphable Models (3DMMs) have become the standard for representing facial geometry, recovering high-fidelity, animatable portraits from monocular images remains an ill-posed problem. Current approaches predominantly rely on face-centric

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.

models (e.g., FLAME Li et al. (2017)) or full-body models (e.g., SMPL-X Pavlakos et al. (2019)). However, despite significant advances in shape reconstruction accuracy, existing methods still lack a torso-rooted reference for joint head–torso estimation, and they often let expression blendshapes absorb jaw opening. We view this difficulty as a chain of three under-specifications in current portrait recovery pipelines: a representation gap, a supervision gap, and a jaw–expression shortcut gap. The first gap is representational. Most state-of-the-art face reconstruction methods Deng et al. (2019); Feng et al. (2021b); Zielonka et al. (2022) crop the input to the facial region, discarding the neck and torso. Even when their camera parameterization differs, their reconstruction is effectively head-centered because the model lacks a torso-rooted kinematic reference. Without the torso as a kinematic anchor, camera motion and head–neck articulation are not parameterized in a common reference frame. In synthesized animations, the head then rotates around its center rather than the anatomical neck joint. These kinematic inconsistencies also make existing talking-avatar models harder to train and generalize, as motion must be learned without a stable head–torso reference frame. The second gap appears after introducing a more complete portrait representation: the torso, neck, and shoulder variables are much more weakly supervised than the face in in-the-wild images. Sparse keypoints from human foundation models such as Sapiens-2 Khirodkar et al. (2026) provide useful face and shoulder cues, but they can be noisy or missing under cropping, clothing, occlusion, and profile views. Therefore, a portrait model should not rely on a single pseudo-label source; it needs complementary supervision in feature space, projection space, and relative 3D geometry.

The third gap is jaw–expression shortcutting. In the FLAME parameter space, mouth opening can be approximated either by rotating the jaw joint or by activating expression coefficients. Since neural networks prioritize shortcut learning, regressors often over-rely on expression parameters to fit mouth movements. This leaves the jaw bone static during speech. While recent works attempt to mitigate this via soft regularization (e.g., lip-reading losses Filntisis et al. (2023) or emotion consistency Daněček et al. (2022)), jaw rotation and mouth-opening blendshapes can still trade off. On talking sequences, predicted jaw rotation often remains weakly correlated with skeletal mouth opening while expression coefficients absorb most of the motion; we report a skeletal-consistency proxy for this mismatch in Section 4. The resulting meshes can look plausible in a single frame but fail under jaw-driven re-animation.

To address these limitations, we formulate portrait recovery as a progressive resolution of the three gaps above rather than as a flat parameter regression problem. We propose **graduated routing**: features and supervision follow the kinematic order from global pose to local expression. The torso anchors the neck, the neck drives the head, and the jaw articulates before soft tissues deform. Ignoring this order leaves pose and expression coupled.

Based on this insight, we present **GRAPE** (Graduated Routing for Articulated Portrait mesh Estimation). First, we formulate portrait geometry with the **Portrait Parametric Model (PPM)**, a task-specific representation that roots the head–neck–shoulder chain at the torso. PPM reuses a FLAME head and an SMPL-X torso so that joint head–torso estimation shares one kinematic reference. Second, we propose a **Progressive Anatomical Alignment (PAA) network**: a frozen Sapiens-2 encoder Khirodkar et al. (2026) extracts features, three experts regress PPM parameters from coarse to fine, and a **Graduated-Mask Router** randomly drops deeper supervision during training to reduce shortcut learning. Third, to supervise the richer but weakly labeled portrait variables, we use multi-source supervision: Sapiens-2 keypoints for sparse anatomical anchors, Pixel3DMM feature distillation for representation-level geometry priors, foreground projection masks for visible portrait extent, and relative geometry losses for head–torso layout.

In summary, our contributions are:

- We formulate the Portrait Parametric Model (PPM) as a task-specific articulated portrait representation. Its role is a torso-rooted coordinate and kinematic reference for joint head–torso estimation.
- We propose Progressive Anatomical Alignment (PAA) with a Graduated-Mask Router, a training formulation that regresses pose before expression, so jaw pitch stays consistent with skeletal mouth opening and expression is less able to shortcut that motion.
- We introduce multi-source supervision for PPM, combining Sapiens-2 keypoints, Pixel3DMM feature distillation, foreground projection masks, and relative geometry constraints to stabilize weakly supervised portrait variables.

- Experiments show stronger image alignment, higher jaw–skeletal consistency, and less jaw–expression shortcutting, with head/face vertex error comparable to strong face methods, and improved training/reconstruction behavior of downstream talking-head and avatar models.

2 Related Work

Monocular reconstruction and the floating-head ambiguity. Face-centric 3DMMs such as DECA Feng et al. (2021b), MICA Zielonka et al. (2022), and TEASER Liu et al. (2025) recover high-fidelity faces from cropped images, but they discard the neck and torso. GNM Ploumpis et al. (2026) includes the neck, teeth, and tongue, but it has no torso anchor and no dedicated jaw joint, so jaw motion stays in the expression basis. This "floating head" assumption leaves neck articulation and camera pose without a shared torso reference and weakens downstream talking-head models. Full-body methods such as SMPL-X Pavlakos et al. (2019) and PIXIE Feng et al. (2021a) provide a spine-rooted kinematic chain, yet their facial spaces lack FLAME-level expression. PEAR Wu et al. (2026) also recovers a head–shoulder portrait, but the head and torso often remain weakly attached. PPM places a FLAME2023 head in the SMPL-X torso frame, so joint head–torso estimation has one torso-rooted kinematic reference Li et al. (2017; 2023).

Jaw–expression shortcutting. A second line of work targets identity–pose–expression control with soft constraints, including emotion losses in EMOCA Daněček et al. (2022) and lip-reading losses in SPECTRE Filntisis et al. (2023). These regularizers still allow jaw rotation and mouth-opening blendshapes to trade off, so re-animation can leave the jaw nearly static Filntisis et al. (2023); Daněček et al. (2022). Appearance-based scores such as landmark mouth opening can be inflated by expression shortcuts and therefore mis-rank how much of the mouth opening is carried by the jaw. We instead use coarse-to-fine routing during training to encourage articulation to explain the image before expression is available.

Pseudo-labels and foundation priors. In-the-wild 3D labels are scarce, so recent pipelines fit pseudo-labels from strong regressors such as TEASER Liu et al. (2025) and ProHMR Kolotouros et al. (2021). Naive joint fitting inherits the entanglement of the underlying model. Foundation priors such as Pixel3DMM Giebenhain et al. (2025) and Sapiens-2 Khirodkar et al. (2026) are typically attached to a single 3DMM vector. We stage pseudo-label fitting as Global $\rightarrow$ Articulation $\rightarrow$ Details, and route these priors through anatomically ordered experts. More discussion is in Section A.

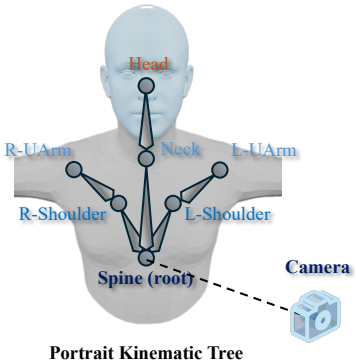
3 Method

Given a portrait image, we estimate PPM parameters and reconstruct an animatable portrait mesh. Specifically, our method addresses three linked gaps. PPM closes the representation gap by giving joint head–torso estimation a torso-rooted kinematic reference. Progressive Anatomical Alignment (PAA) is the matching training formulation: parameters are supervised in the same coarse-to-fine order as the portrait kinematic tree, which targets jaw–skeletal consistency and reduces expression shortcuts for mouth opening. Multi-source and ordinal supervision closes the supervision gap by combining sparse anatomical keypoints with feature-, mask-, and geometry-level constraints. Section 3.1 defines PPM; Section 3.2 describes PAA; Section 3.3 summarizes data and training.

3.1 Portrait Parametric Model

Most 3DMM face models treat head pose as a single global rotation. This is sufficient for cropped face alignment but under-specifies a portrait, where the visible neck and shoulders define the natural reference frame for head motion. Physically, head pose is the rotation of the neck relative to the torso, while camera pose is an independent extrinsic variable. Without a torso anchor, these quantities lack a shared reference frame, which leads to the "floating head" problem in monocular reconstruction and animation.

Full-body models such as SMPL-X provide a torso anchor, but their built-in face space is limited: the body shape space has only 10 dimensions, and the facial expression space is too low-dimensional to capture fine talking-face motion. We therefore instantiate PPM as a task-specific portrait representation $\mathcal{M}_{\mathrm{PPM}}$: FLAME2023 supplies head shape and expression Li et al. (2017; 2023), and the



Portrait Kinematic Tree

Algorithm-1: Canonical Space Injection for Mesh Fusion (PPM forward)

Input: Head shape β, Expression ψ, Body shape α, Jaw & Eyes θ_h, Torso θ_t
Output: Portrait 3D Vertices (V_{out}, F)

1. $V_{head} = FLAME23(\beta, \phi, \theta_h)$ // Get head with expressions
2. $V_{portrait} = SMPL_T(\alpha, \mathbf{0})$ // Get canonical portrait torso
3. Calculate the center of two eyes of $V_{head} \rightarrow C_h$
4. Calculate the center of two eyes of $V_{portrait} \rightarrow C_p$
5. Replace head in $V_{portrait}$ by $V_{head} - C_h + C_p \rightarrow V^*_{portrait}$ Canonical space
6. Smoothing neck boundary vertices
7. $V_{out} = LBS(V^*_{portrait}, \theta_t)$
Return (V_{out}, F) // F is the mesh faces

Canonical Injection

Figure 2: **Portrait Parametric Model (PPM). Left:** Kinematic tree rooted at the spine; the torso anchors camera-relative pose. **Right:** Canonical Injection. The FLAME head and SMPL-T are aligned in canonical space via eye centers ($C_h \rightarrow C_p$) and merged into one watertight mesh.

SMPL-X torso (termed SMPL-T) supplies the spine-rooted frame for neck and shoulder kinematics Pavlakos et al. (2019). Specifically, we remove the lower-body and arm meshes and update the SMPL-X LBS weights to simplify the torso representation.

3.1.1 Kinematic-aware Topology

PPM maps portrait parameters to a mesh $\mathbf{P} = \{\mathbf{V}, \mathbf{F}\}$:

$$\mathbf{P} = \mathrm{PPM}(\boldsymbol{\beta}_h, \boldsymbol{\beta}_t, \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\theta}), \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{N_v \times 3}$ and $\mathbf{F} \in \mathbb{N}^{N_f \times 3}$ are vertices and faces. Head shape $\boldsymbol{\beta}_h$ and expression $\boldsymbol{\psi}$ come from FLAME2023 Li et al. (2023); torso shape $\boldsymbol{\beta}_t$ comes from SMPL-X Pavlakos et al. (2019); $\mathbf{s}$ is a global scale of the head mesh; pose $\boldsymbol{\theta}$ covers the spine (root), neck, head, jaw, and eyes. Shoulder and upper-arm poses $\boldsymbol{\theta}_{\mathrm{sh}}$ and $\boldsymbol{\theta}_{\mathrm{ua}}$ are included for torso movement. As shown in Fig. 2 (Left), the forward kinematic tree for the head is

$$\text{Spine (Root)} \rightarrow \text{Neck} \rightarrow \text{Head} \rightarrow \{\text{Jaw, Eyes}\}. \tag{2}$$

The camera is defined relative to the spine root rather than the head center. Because the head joint is a child of the neck ($\mathbf{J}_{\mathrm{head}} \in \mathrm{child}(\mathbf{J}_{\mathrm{neck}})$), head rotation is expressed relative to the neck. This reduces the ambiguity between intrinsic articulation and extrinsic camera motion.

3.1.2 Canonical Injection Strategy

Replacing the SMPL-X head with FLAME improves facial fidelity, but stitching the two meshes in posed space is unstable: small pose errors cause gaps or self-intersections at the neck ring. We instead fuse them in zero-pose canonical space, where alignment depends only on shape (Fig. 2, Right). Concretely, we (i) build a canonical FLAME head with jaw and expression only, (ii) build an SMPL-T torso, (iii) align the two by eye-socket centers C_h and C_p, and (iv) replace the SMPL-X head, apply neck-ring smoothing, and drive the unified mesh with LBS. Construction, neck-ring smoothing, and skinning are in Section B.1.

3.2 Progressive Anatomical Alignment

Given PPM, we estimate its parameters with **Progressive Anatomical Alignment (PAA)**. PPM exposes the correct portrait variables, but directly regressing all of them remains under-constrained: shoulder keypoints may be noisy, torso geometry is weakly observed, and jaw pose can trade off against expression. The key idea is to match the anatomical ordering of human motion: global shape and camera first, then skeletal articulation, then non-rigid expression. PAA implements this ordering through both the network architecture and the loss design.

As shown in Fig. 3, the PAA network has three stages:

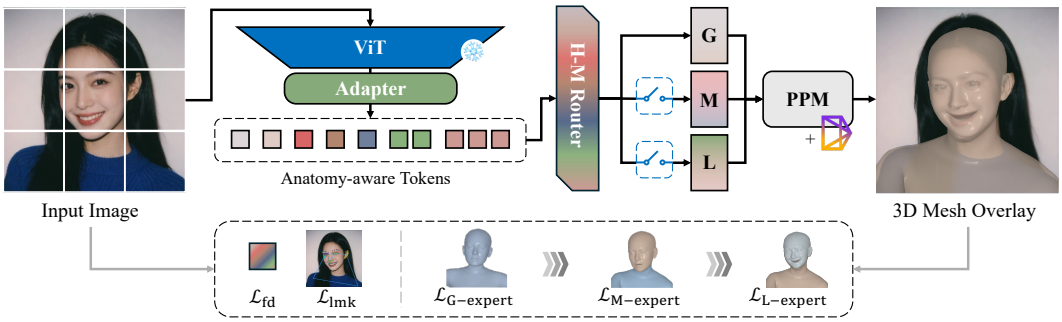


Figure 3: **Progressive Anatomical Alignment (PAA).** A frozen pretrained encoder extracts image features; a learnable adapter maps them to anatomy-aware tokens. Each expert has a learnable query embedding that cross-attends to these tokens (keys/values) to predict the corresponding PPM subset. The Graduated-Mask Router gates M- and L-expert losses during training.

1. A pretrained ViT-based feature extractor extracts portrait features from the input image.
2. A learnable adapter converts encoder features into anatomy-aware tokens. Each expert uses its own learnable embedding as the query and these tokens as keys/values in cross-attention, then an MLP regresses the corresponding body-part PPM parameters.
3. The predicted parameters are passed through PPM to obtain $\mathbf{V}_{\text{out}}$, which is projected to the image plane for 2D/3D supervision.

3.2.1 Anatomy-aware Token Extraction

Since a global CLS token mixes shape, pose, and expression in one representation, which makes expression shortcuts easier for downstream regressors, we use a query-based Transformer decoder rather than a single global token. The input image $\mathbf{I}$ is encoded by a frozen Sapiens-2 encoder Khirodkar et al. (2026) into spatial features $\mathbf{Z}_{\text{img}} \in \mathbb{R}^{N\times D}$. Sapiens-2 is pretrained on large-scale human images and provides strong portrait features. We then apply a learnable feature adapter to obtain anatomy-aware tokens $\mathbf{T}_{\text{anat}} \in \mathbb{R}^{N\times D}$, which serve as the keys and values for expert-wise cross-attention. For each expert level $k \in \{G, M, L\}$, we introduce a dedicated learnable query embedding $\mathbf{q}_k$. The expert aggregates the corresponding body-part cues by attending to $\mathbf{T}_{\text{anat}}$:

$$\mathbf{z}_k = \text{CrossAttn}(\mathbf{q}_k, \mathbf{T}_{\text{anat}}), \quad k \in \{G, M, L\}. \tag{3}$$

$\mathbf{z}_G$ aggregates information for global shape and camera; $\mathbf{z}_M$ focuses on neck, head, and jaw joints , as well as the shoulders and upper arms; $\mathbf{z}_L$ focuses on expression and eye details. This splits the regression problem at the feature level before any parameter is predicted.

3.2.2 Graduated-Mask Hierarchical Experts

The attended feature $\mathbf{z}_k$ is passed to a dedicated MLP expert that predicts the PPM parameters of the corresponding anatomical part: 1) **G-Expert ($\mathbf{z}_G$)** predicts camera $\boldsymbol{\pi}$, head shape $\boldsymbol{\beta}_h$, torso shape $\boldsymbol{\beta}_t$, and global scale $\mathbf{s}$. 2) **M-Expert ($\mathbf{z}_M$)** predicts $\boldsymbol{\theta}_{\text{spine}}$, $\boldsymbol{\theta}_{\text{neck}}$, $\boldsymbol{\theta}_{\text{head}}$, $\boldsymbol{\theta}_{\text{jaw}}$, $\boldsymbol{\theta}_{\text{sh}}$, and $\boldsymbol{\theta}_{\text{ua}}$. 3) **L-Expert ($\mathbf{z}_L$):** expression $\boldsymbol{\psi}$, eye gaze $\boldsymbol{\theta}_{\text{eye}}$. The predicted parameters are fed into PPM to produce $\mathbf{V}_{\text{out}}$, and the mesh is projected with $\boldsymbol{\pi}$ for loss computation.

Graduated-Mask Router. During training, we randomly disable supervision on deeper experts so that each level must remain useful on its own. Let $m_k \in \{0, 1\}$ be the mask for level k, with $m_G = 1$ always:

$$m_M = m_G \cdot b_M, \quad b_M \sim \text{Bernoulli}(p), \tag{4}$$

$$m_L = m_M \cdot b_L, \quad b_L \sim \text{Bernoulli}(p). \tag{5}$$

When $m_L = 0$ but $m_M = 1$, the forward pass sets $\boldsymbol{\psi} = \mathbf{0}$ and the network must explain the image with shape and articulation alone; expression cannot absorb jaw errors. $\mathcal{L}_{\text{lmk}}$ remains active. This discourages the shortcut where expression blendshapes mimic mouth opening. At test time all masks are 1; the expression expert then fits the residual left by articulation.

3.2.3 TRAINING OBJECTIVES

In-the-wild images lack 3D ground truth. We therefore train PAA with multi-source supervision. Sapiens-2 Khirodkar et al. (2026) and 68-point landmarks Zhu et al. (2016) provide sparse 2D anatomical anchors; Pixel3DMM Giebenhain et al. (2025) feature distillation transfers representation-level 3D priors; the overflow mask constrains the projected torso extent; and relative geometry preserves the head–torso layout when shoulder detections are missing or noisy. We fit pseudo-labels offline with the Graduated-Refinement Pipeline (Section C) and optimize

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{ord}} + \mathcal{L}_{\text{fd}} + \underbrace{\mathcal{L}_{\text{lmk}} + \mathcal{L}_{\text{reg}}^{G}}_{\text{G-Expert}} + m_M \underbrace{\left(\mathcal{L}_{\text{nc}} + \mathcal{L}_{\text{om}} + \mathcal{L}_{\text{rg}} + \mathcal{L}_{\text{reg}}^{M}\right)}_{\text{M-Expert}} \\ + m_L \underbrace{\left(\mathcal{L}_{\text{lmk}}^{\text{mouth}} + \mathcal{L}_{\text{lmk}}^{\text{eye}} + \mathcal{L}_{\text{lmk}}^{\text{pupil}} + \mathcal{L}_{\text{reg}}^{L}\right)}_{\text{L-Expert}}.\end{aligned} \tag{6}$$

Mask-gated terms follow the expert that supervises them; $\mathcal{L}_{\text{lmk}}$ and $\mathcal{L}_{\text{fd}}$ stay active. $\mathcal{L}_{\text{ord}}$ is L2 regression to pseudo-label parameters; $\mathcal{L}_{\text{fd}}$ distills Pixel3DMM features; $\mathcal{L}_{\text{lmk}}$ anchors global 2D landmarks; $\mathcal{L}_{\text{nc}}$ matches the nose–chin opening distance; $\mathcal{L}_{\text{om}}$ penalizes torso overflow outside the foreground; $\mathcal{L}_{\text{rg}}$ preserves relative head–torso geometry; and the remaining terms supervise mouth/eye/pupil landmarks with parameter regularization. Full definitions of $\mathcal{L}_{\text{ord}}$, $\mathcal{L}_{\text{fd}}$, $\mathcal{L}_{\text{lmk}}$, $\mathcal{L}_{\text{nc}}$, $\mathcal{L}_{\text{om}}$, $\mathcal{L}_{\text{rg}}$ and the detail-landmark terms are in Section D. Overflow-mask examples are in Fig. 8.

3.3 IMPLEMENTATION DETAILS

Data preparation. We fit PPM parameters offline as pseudo-labels (details are in Section C). For each training image, we also extract 68-point landmarks Zhu et al. (2016), Sapiens-2 keypoints (face and shoulder), a foreground segmentation mask, and SMPL-X body parameters. These signals supervise different expert levels in Eq. (6).

Two-phase training. Training has two stages aligned with the graduated design. In stage 1, we pretrain on the multi-view Nersemble dataset Kirschstein et al. (2023) to learn stable shape and camera estimation; both m_M and m_L are always 1. In stage 2, we finetune on LS3DW Bulat & Tzimiropoulos (2017), CelebA Liu et al. (2015), LaPa Liu et al. (2020), and LFW Huang et al. (2007) with graduated masking enabled ($p = 0.5$) so that the router improves jaw–skeletal consistency and reduces jaw–expression shortcutting on in-the-wild data. All experiments use 4 NVIDIA A100 GPUs, Adam optimizer, batch size 64, and 4M iterations ($\sim$5 days). Additional hyperparameters are in the appendix (Implementation Details).

4 EXPERIMENTS

We organize the experiments around five questions. **RQ1:** reconstruction quality. **RQ2:** can GRAPE improve jaw–skeletal consistency and reduce jaw–expression shortcutting? **RQ3:** head–shoulder kinematics. **RQ4:** each component's contribution. **RQ5:** downstream applications.

4.1 EXPERIMENTAL SETUP

We follow the two-phase training in Section 3.3, with PPM pseudo-labels from the Graduated-Refinement Pipeline (Section C). We evaluate (1) image landmarks on LS3DW Bulat & Tzimiropoulos (2017), CelebA Liu et al. (2015), LaPa Liu et al. (2020), and LFW Huang et al. (2007); (2) talking-head video on 20 randomly selected HDTF Zhang et al. (2021) clips; and (3) head/face vertex error on NersembleV2 Kirschstein et al. (2023), NoW Sanyal et al. (2019), and Stirling Feng et al. (2018). We report 68-point NME (excluding jawline indices 0–16) Wu et al. (2026), JSR/temporal scores on video, and head/face MVE/LVE Feng et al. (2021b) plus NoW/Stirling mean/median/std. Quantitative baselines include DECA Feng et al. (2021b), EMOCA Daněček et al. (2022), Deep3DFace Deng et al. (2019), TEASER Liu et al. (2025), SPECTRE Filntisis et al. (2023), 3DDFA-V2 Guo et al. (2020), 3DDFA-V3 Wang et al. (2024), SMPLest-X Yin et al. (2026), and PEAR Wu et al. (2026), each with its official crop. SAM-3DB Yang et al. (2026) is a qualitative reference in Figs. 1 and 4 and is excluded from the tables. All reported metrics are computed on held-out test images that do not overlap the training set. Metric definitions are in Section E.

4.2 Qualitative Comparison

Fig. 4 compares mesh overlays on in-the-wild portraits with visible shoulders. FLAME/BFM-based baselines often detach the head from the torso, while GRAPE recovers a continuous head–neck–shoulder mesh. PEAR and SAM-3DB can predict plausible shoulders, but PEAR often leaves a head–torso gap and SAM-3DB does not estimate expression. This supports **RQ1** and **RQ3**.

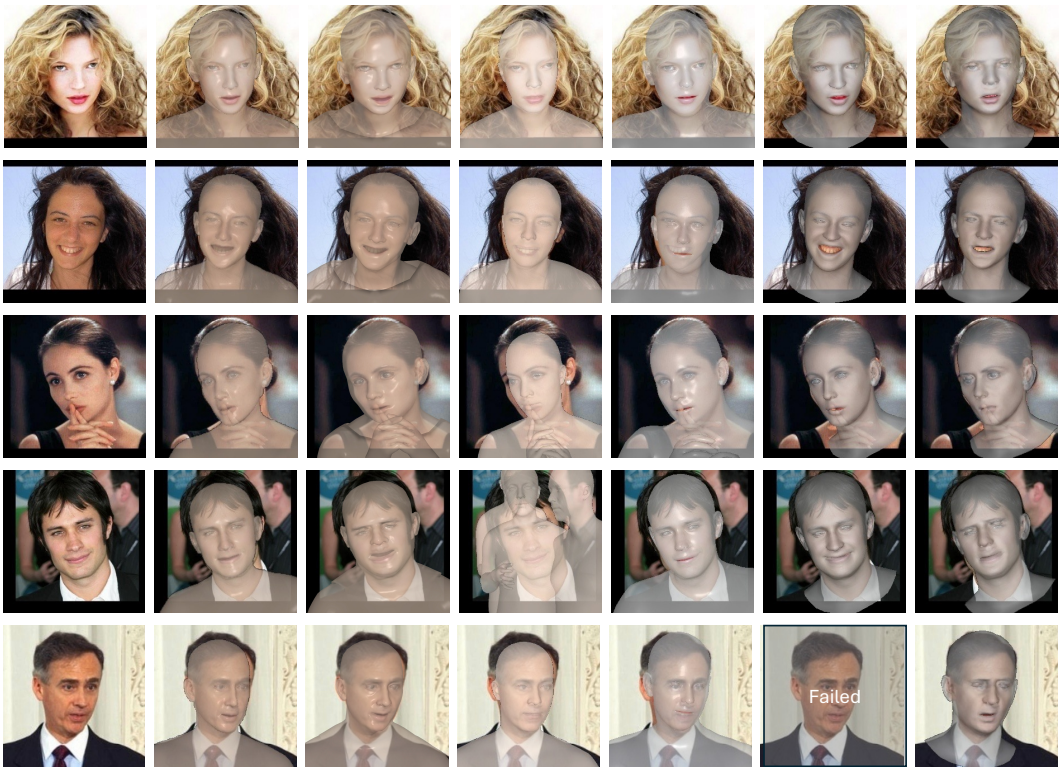


Figure 4: **Qualitative reconstruction on in-the-wild images.**

To answer **RQ2**, we compare predicted jaw pitch (θ_{j-p}) with a skeletal-consistency proxy during speech. JCR is a normalized nose–chin distance, and JSR is the Pearson correlation between jaw pitch and JCR (defined in Section E.2). JCR uses the same observed nose–chin landmarks as $\mathcal{L}_{\mathrm{nc}}$, so a higher JSR means that jaw pitch tracks this 2D cue. We use that agreement, together with the with/without-expression frames, to assess jaw–skeletal consistency and jaw–expression shortcutting. Fig. 5 shows a case where TEASER Liu et al. (2025) attributes a smile to expression while the jaw stays inconsistent with that cue; GRAPE tracks JCR more closely. Additional frame-level comparisons with PEAR, SPECTRE, and TEASER are in Fig. 9.

4.3 Quantitative Comparison

(1) Image landmark accuracy. Table 1 reports 68-point NME on four in-the-wild datasets. GRAPE has the lowest average, ahead of TEASER, which does not model the torso. SPECTRE and SMPLest-X are much higher because their predicted landmarks are poorly aligned to the face. **(2) Talking-head motion.** On HDTF (Table 2), GRAPE obtains the highest skeletal-consistency proxy (JSR) and the best face and shoulder NME. Temporal stability is second to SMPLest-X, whose 10-D expression space produces smaller mouth motion. **(3) Head/face vertex accuracy.** Table 3 scores head and face vertices only: the shared FLAME head on NersembleV2, and face scan-to-mesh distance on NoW and Stirling. GRAPE stays within a few hundredths of TEASER Liu et al. (2025) on NoW and Stirling, so the neck–shoulder extension preserves facial surface accuracy. The lower NersembleV2 error covers only the FLAME head after pretraining on that dataset.

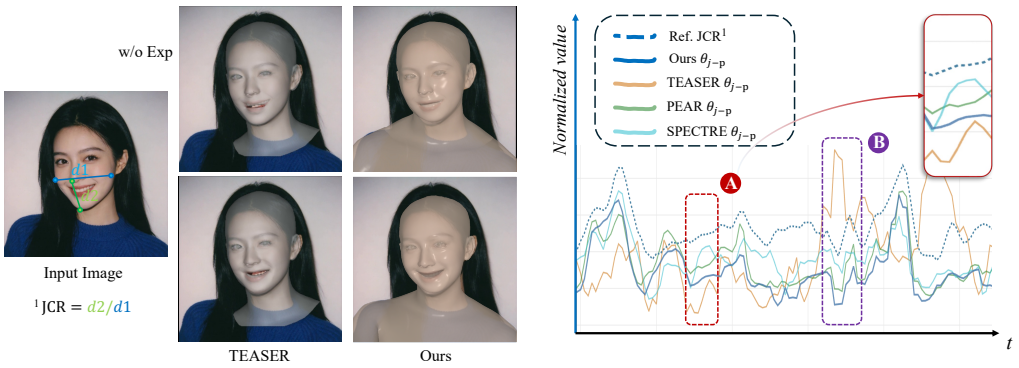


Figure 5: **Jaw–skeletal consistency during speech. Left:** reconstructions with and without expression on the same talking frame. **Right:** time series of the normalized nose–chin reference JCR (blue) versus predicted jaw pitch for GRAPE (beige) and TEASER (black). The highlighted interval shows TEASER attributing mouth motion to expression while jaw pitch fails to track JCR.

Table 1: Projection alignment comparison in image space. We report 68-keypoints-based normalized mean error (NME $\times 100$) excluding the jawline keypoints. Each method uses its own official crop. Large NME reflects poor alignment of the predicted facial landmarks.

Method	3DMM Basis	LS3DW ↓	CelebA ↓	LaPa ↓	LFW ↓	Avg. ↓
DECA Feng et al. (2021b)	FLAME2020	23.88	21.60	25.99	20.01	22.87
EMOCA Daněček et al. (2022)	FLAME2020	31.03	22.49	29.99	19.98	25.87
SPECTRE Filntisis et al. (2023)	FLAME2020	38.51	36.32	31.21	35.70	35.44
3DDFA-V3 Wang et al. (2024)	BFM	18.42	13.66	15.95	13.65	15.42
TEASER Liu et al. (2025)	FLAME2020	<u>15.72</u>	<u>7.01</u>	<u>9.88</u>	10.26	<u>10.72</u>
SMPLest-X Yin et al. (2026)	SMPL-X	51.79	49.99	41.53	40.29	45.90
PEAR Wu et al. (2026)	EHM-s	19.24	8.74	11.87	<u>10.02</u>	12.47
GRAPE (Ours)	PPM	**10.56**	**5.63**	**7.07**	**8.22**	**7.87**

4.4 Ablation Studies and Analysis

We remove one component at a time (Table 4), answering **RQ4**. A flat regressor drops the anatomy-aware queries and predicts every PPM parameter from one pooled feature. The other rows disable the router, train on jointly fitted pseudo-labels, or replace PPM with a floating-head FLAME. Removing PAA drops JSR from 0.799 to 0.368 and raises NME to 10.44, so it accounts for most of the JSR gain. The router adds a smaller JSR lift, from 0.710 to 0.799 (V1.5 in Table 5), while NME and mesh error stay close. Removing graduated pseudo-labels raises NME to 12.55 while JSR stays at 0.701, so the labels carry more of the alignment gain. The torso anchor shows up in MPJPE, which includes shoulder joints; without it MPJPE rises from 69.4 to 89.1, while face-only NME and mesh error stay close. Loss and router sweeps are in Table 5, Fig. 10, and Section E.

Table 4: Structural ablation of the proposed anatomical design. Each row removes one high-level component while keeping the remaining training recipe unchanged. JSR is the skeletal-consistency proxy on HDTF; NME ($\times 100$) measures image-space face alignment; MVE ($\times 10^3$)/LVE ($\times 10^4$) measure 3D head/mouth mesh accuracy; MPJPE measures head–shoulder joint accuracy.

Variant	HDTF JSR ↑	Avg. NME ↓	Ner-MVE ↓	Ner-LVE ↓	MPJPE ↓
Flat regressor w/o PAA	0.368	10.44	24.89	26.01	79.0
w/o Graduated-Mask Router	0.710	<u>8.07</u>	8.31	6.20	<u>71.3</u>
w/o graduated pseudo-labels	0.701	12.55	11.92	9.53	76.6
w/o PPM torso anchor	<u>0.782</u>	8.11	<u>8.23</u>	**6.15**	89.1
GRAPE full	**0.799**	**7.87**	**8.20**	<u>6.19</u>	**69.4**

Table 2: Facial motion evaluation on the HDTF test set (20 videos). JSR is the skeletal-consistency proxy for RQ2: Pearson r between jaw pitch and the nose–chin reference JCR (Fig. 5), which reuses the observed landmarks supervised by $\mathcal{L}_{\text{nc}}$. Var-Exp / Var-Jaw are temporal variances; values that are too large or too small indicate unbalanced dynamics. Temporal is the mean second-order difference (↓ smoother). Color coding of extreme variances is illustrated in Fig. 9.

Method	JSR ↑	(Var-Exp	Var-Jaw)	Temporal ↓	NME-P ↓	NME-S ↓
SPECTRE Filntisis et al. (2023)	0.564	0.104	0.601	1.847	0.179	–
TEASER Liu et al. (2025)	0.371	0.490	0.197	2.344	0.066	–
SMPLest-X Yin et al. (2026)	0.591	0.003	0.207	**1.040**	0.117	0.260
PEAR Wu et al. (2026)	0.537	0.414	0.123	1.851	0.072	0.849
GRAPE (Ours)	**0.799**	0.297	0.348	1.680	**0.058**	**0.155**

Table 3: Head/face vertex error. Nersemble-V2 reports MVE ($\times 10^3$) and LVE ($\times 10^4$) on the shared FLAME head; NoW and Stirling report face scan-to-mesh distance.

Method	Nersemble-V2		NoW benchmark			Stirling benchmark		
	MVE ↓	LVE ↓	Median ↓	Mean ↓	Std ↓	Median ↓	Mean ↓	Std ↓
Deep3DFace Deng et al. (2019)	–	–	1.11	1.41	1.21	**0.99**	1.27	1.15
DECA Feng et al. (2021b)	60.67	74.72	1.23	1.57	1.39	1.03	1.32	1.18
EMOCA Daněček et al. (2022)	63.29	36.37	1.24	1.56	1.44	1.02	1.32	1.28
3DDFA-V2 Guo et al. (2020)	–	–	1.09	1.38	1.18	1.20	1.55	1.45
3DDFA-V3 Wang et al. (2024)	–	–	1.05	1.33	1.20	1.19	1.56	1.43
SPECTRE Filntisis et al. (2023)	67.92	59.32	1.36	1.37	1.27	1.04	1.19	1.25
SMIRK Retsinas et al. (2024)	37.88	40.19	0.99	1.22	1.02	1.01	1.08	1.05
TEASER Liu et al. (2025)	35.04	32.02	**0.92**	1.10	**0.99**	1.00	**1.07**	1.04
PEAR Wu et al. (2026)	13.62	8.08	0.94	1.17	1.08	1.03	1.20	1.11
GRAPE (Ours)	**8.20**	**6.19**	0.93	**1.09**	**0.99**	**0.99**	1.08	**1.03**

4.5 Applications

Audio-driven talking-head generation. To test **RQ5**, we train DiffPoseTalk Sun et al. (2024) on the same 10 HDTF clips twice, once with TEASER parameters and once with GRAPE parameters. The GRAPE-labeled set shows a modest training-loss decrease (Fig. 11). We read this, together with the RGBAvatar reconstruction results below, as improved training/reconstruction behavior of the downstream models.

Animatable 3D Gaussian avatar generation. We train RGBAvatar Li et al. (2025) on the same HDTF sequences with TEASER parameters, GRAPE head-only parameters, and full PPM. Full PPM reduces the floating-head artifact at the neck and shoulders (Fig. 12). Full-frame SSIM/PSNR rise slightly over both (Table 6), and body-region PSNR also improves over the head-only model.

5 Conclusion

We presented **GRAPE** for monocular portrait mesh estimation. PPM is a task-specific portrait representation: rooting the kinematic chain at the torso, with the FLAME head attached to the SMPL-X neck in canonical pose, gives joint head–torso estimation a shared reference and reduces the ambiguity between intrinsic articulation and extrinsic camera motion. The training formulation predicts jaw pose before expression, which improves jaw–skeletal consistency and reduces expression shortcuts for mouth opening. Image alignment and this skeletal-consistency proxy improve, while face vertex error stays comparable, and the parameters improve the training/reconstruction behavior of downstream talking-head and avatar models.

Limitations and Future Work. The neck and shoulders still assume a tight skin surface, so loose clothing and heavy neck occlusion remain failure cases; the face region is more stable. Pseudo-label quality follows the 2D landmarks, which are noisy under extreme lighting and profile views. Future work should keep this jaw-level control while representing hair and clothing that the skin mesh cannot.

AI USE STATEMENT

We used generative image models to create the input portraits in Fig. 1. These images are illustration-only figures. We also used generative AI tools to polish wording and to assist with code editing. We have not used generative AI tools to generate training datasets, design the method, write proofs, or interpret experimental results. All AI-assisted text, code, and teaser images were reviewed by the authors, who take responsibility for the final content of this work, including claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Portrait reconstruction and animation can support telepresence, accessibility, digital avatars, and creative tools, but they can also be misused for impersonation or non-consensual manipulation. This work studies geometric mesh recovery and torso-rooted portrait kinematics, not identity transfer or photorealistic face synthesis. Training and evaluation use public datasets (NersembleV2, LS3DW, CelebA, LaPa, LFW, HDTF, NoW, and Stirling) under their respective licenses; we do not release new biometric datasets. The AI-generated portraits in Fig. 1 are synthetic illustrations, not photographs of real people, and are not used as training data. Applications built on GRAPE should obtain consent, disclose generated or animated content when appropriate, and avoid deceptive or unauthorized biometric uses.

REPRODUCIBILITY STATEMENT

Architecture, losses, and the two-phase training recipe are specified in Sections 3 and 3.3, with full loss definitions in Section D and hyperparameters in Section B. PPM construction, including canonical injection and neck-ring smoothing, is in Section B.1. The offline Graduated-Refinement Pipeline for pseudo-labels is in Section C. Dataset splits, metrics, and evaluation protocols are in Sections 4.1 and E. We include anonymized source code with the supplementary material.

REFERENCES

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1021–1030, 2017. doi: 10.1109/ICCV.2017.116.

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 (CVPR)*, pp. 20311–20322, 2022.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 285–295, 2019.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pp. 792–804. IEEE, 2021a.

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, 2021b.

Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 780–786. IEEE, 2018.

Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. SPECTRE: Visual speech-informed perceptual 3D facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5745–5755, 2023.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *International Conference on Learning Representations (ICLR)*, 2026.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 42(4): 161:1–161:14, 2023. doi: 10.1145/3592455.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2252–2261, 2019.

Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11605–11614, 2021.

Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. RGBAvatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10747–10757, 2025.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194:1–194:17, 2017.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. FLAME: Learning a model of facial shape and expression from 4d scans (2023 release). `https://flame.is.tue.mpg.de/`, 2023.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21159–21168, 2023.

Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11637–11644, 2020.

Yunfei Liu, Lei Zhu, Lijian Lin, Ye Zhu, Ailing Zhang, and Yu Li. Teaser: Token enhanced spatial modeling for expressions reconstruction. In *ICLR*, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Stylianos Ploumpis, Jan Bednarik, Gaspard Zoss, Ruslan Guseinov, Luca Prasso, Prashanth Chandran, et al. GNM head: A generative aNthropometric model of the human head. *arXiv preprint arXiv:2607.23687*, 2026.

Zesong Qiu, Yuwei Li, Dongming He, Qixuan Zhang, Longwen Zhang, Yinghao Zhang, Jingya Wang, Lan Xu, Xudong Wang, Yuyao Zhang, and Jingyi Yu. SCULPTOR: Skeleton-consistent face creation using a learned parametric generator. *ACM Transactions on Graphics (TOG)*, 41(6), 2022. doi: 10.1145/3550454.3555462.

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 (CVPR)*, pp. 2490–2501, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7763–7772, 2019.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-Jin Liu. DiffPoseTalk: Speech-driven stylistic 3D facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4), 2024. doi: 10.1145/3658221.

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1274–1283, 2017.

Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, and Zhen Lei. 3d face reconstruction with the geometric guidance of facial part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1672–1682, 2024.

Jiahao Wu, Yunfei Liu, Lijian Lin, Ye Zhu, Lei Zhu, Jingyi Li, and Yu Li. PEAR: Pixel-aligned expressive human mesh recovery. In *ACM SIGGRAPH 2026 Conference Papers*, 2026.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. SAM 3D body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. SMPLest-X: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1778–1794, 2026. doi: 10.1109/TPAMI.2025.3618174.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3661–3670, 2021.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Haiming Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 146–155, 2016.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, pp. 250–269. Springer, 2022.

Appendix

This appendix contains:

- **Extended Related Work.** Additional discussion of face/body 3DMMs, jaw–expression shortcuts, pseudo-label fitting, and foundation priors.
- **Implementation details.** Hyperparameters, backbones, and expert-net settings.
- **PPM construction.** Canonical injection, neck-ring smoothing, and skinning.
- **Data Generation Pipeline.** Offline Graduated-Refinement for pseudo-labels.
- **Training Objectives.** Full loss definitions corresponding to Eq. (6).
- **Additional Experimental Results.** Extra qualitative results, attention maps, metrics, and downstream applications.
- **Limitations, Discussions and Future Work.** Extended discussion beyond the main-text conclusion.

A Extended Related Work

Face-centric models and internal anatomy. DECA Feng et al. (2021b) uses displacement maps for high-frequency surface detail, while MICA Zielonka et al. (2022) targets metric shape recovery with recognition networks. TEASER Liu et al. (2025) improves expression fidelity with token-based spatial modeling, but still operates on tightly cropped faces. GNM Ploumpis et al. (2026) provides an expressive head model with internal anatomy such as teeth and tongue. However, its jaw motion remains entangled with expression, and the camera is still centered at the head. These face-centric approaches Deng et al. (2019); Feng et al. (2021b); Zielonka et al. (2022); Liu et al. (2025) therefore leave neck articulation and camera pose without a shared torso reference.

Jaw–expression shortcuts. In FLAME-based optimization Li et al. (2017) and regression Daněček et al. (2022), jaw rotation ($\boldsymbol{\theta}_{jaw}$) and mouth-opening blendshapes ($\boldsymbol{\psi}$) are often coupled. Neural regressors frequently converge to a local optimum where expressions compensate for static jaw poses. Appearance-based scores such as landmark mouth opening can be inflated by expression shortcuts and therefore mis-rank how much of the mouth opening is carried by the jaw. The Graduated-Mask Router differs from emotion or lip-reading regularizers by encouraging the network to explain the image with shape and articulation before expression is available. Our JSR score tracks a supervised 2D cue and is the measure we use for jaw–skeletal consistency.

Pseudo-label fitting. Self-supervised photometric reconstruction is standard when in-the-wild 3D ground truth is scarce Tewari et al. (2017); Deng et al. (2019); Sanyal et al. (2019), but photometric losses mix texture, lighting, and depth-scale ambiguity. Recent pipelines generate pseudo-labels by optimization-based fitting Feng et al. (2021b); Lin et al. (2023); Qiu et al. (2022), often warm-started from TEASER Liu et al. (2025) for the face and ProHMR Kolotouros et al. (2021) for the body. Naive joint fitting inherits the entanglement of the underlying model (e.g., fitting a smile with jaw rotation). Our Graduated-Refinement Pipeline stages optimization as Global $\rightarrow$ Articulation $\rightarrow$ Details so that jaw and expression are fit in separate stages (Section C).

Foundation priors as flat heads. Pixel3DMM Giebenhain et al. (2025) distills screen-space geometric cues into face regressors, and Sapiens-2 Khirodkar et al. (2026) provides dense human keypoints. Most prior work applies such features through a flat regression head over a single 3DMM parameter vector. Progressive Anatomical Alignment instead routes frozen Sapiens-2 features and Pixel3DMM distillation through anatomically ordered experts, so representation-level priors attach to shape, articulation, and expression at different stages rather than being mixed in one prediction.

B Implementation Details

This section complements Section 3.3. Portrait crops are resized to 256×256. The cross-attention decoder has 4 layers with hidden size D=768. The learning rate is 10^{-4}. Loss weights are $\mathcal{L}_{\text{ord}}$:

0.1, $\mathcal{L}_{\text{lmk}}$: 2.0, $\mathcal{L}_{\text{nc}}$: 2.0, $\mathcal{L}_{\text{om}}$: 2.5, $\mathcal{L}_{\text{rg}}$: 0.1, and $\mathcal{L}_{\text{fd}}$: 1.0. Exact values are also in the anonymized supplementary code. During NersembleV2 Kirschstein et al. (2023) pretraining, we set $m_M = m_L = 1$ (equivalently p=1.0); during in-the-wild finetuning, the Graduated-Mask Router uses p=0.5. We train with Adam, batch size 64, for 4M iterations on 4 NVIDIA A100 GPUs.

B.1 PPM Construction Details

PPM is a task-specific articulated portrait representation. Its role is a torso-rooted kinematic reference for joint head–torso estimation. It is implemented as a single skinned portrait template whose face region follows FLAME2023 and whose neck–shoulder region follows the upper-body part of SMPL-X. The construction is performed once in canonical space and reused during both pseudo-label fitting and network training.

Torso extraction. Starting from the SMPL-X template, we keep the vertices and faces corresponding to the upper torso, neck, shoulders, and a short upper-arm boundary. Lower-body and hand regions are removed to reduce unnecessary degrees of freedom for portrait images. The remaining torso keeps the SMPL-X shape blendshapes and the spine/neck/shoulder joints used by the portrait kinematic tree.

Four-step canonical injection. The fusion corresponding to Section 3.1 proceeds as follows:

1. **Head mesh.** Build $V_{\text{head}}(\boldsymbol{\beta}_h, \boldsymbol{\psi}, \boldsymbol{\theta}_{\text{jaw}})$ in canonical space. Only jaw articulation and expression are applied; global head rotation is zero.
2. **Torso mesh.** Build $V_{\text{torso}} = \mathbf{J}(\boldsymbol{\beta}_t, \mathbf{0})$ with the SMPL-X torso blendshape function $\mathbf{J}$.
3. **Eye-based alignment.** Compute eye-socket centers C_h (head) and C_p (torso) and translate the head:
$$V^*_{\text{head}} = V_{\text{head}} - C_h + C_p. \tag{7}$$
The FLAME eye-center C_h uses the left/right eyeball or eye-socket landmarks; C_p uses the corresponding SMPL-X head landmarks before head removal. This preserves face scale in the SMPL-X camera-relative coordinate system.
4. **Injection and skinning.** Replace the SMPL-X head vertices with V^*_{head}, apply neck-ring smoothing $\mathcal{S}(\cdot)$, and drive the unified mesh with LBS:
$$V_{\text{out}} = \text{LBS}(\mathcal{S}(V^*_{\text{head}} \cup V^*_{\text{torso}}),\, \boldsymbol{\theta},\, \mathcal{W}_{\text{unified}})\,. \tag{8}$$

Neck-ring smoothing and topology. Let $\mathcal{R}$ denote the fixed set of neck-ring vertices shared by the head and torso templates. For each $i \in \mathcal{R}$, we linearly blend the head and torso positions:

$$\mathcal{S}(v_i) = (1-\alpha_i)\, v_i^{\text{head}} + \alpha_i\, v_i^{\text{torso}}, \quad \alpha_i \in [0,1], \tag{9}$$

where α_i increases monotonically from the head side to the torso side of the ring. Equivalently, in template coordinates, $v_i = (1-\alpha_i)v_i^{\text{FLAME}} + \alpha_i v_i^{\text{SMPL-T}}$. The original SMPL-X head faces are removed above the neck connection band, and the aligned FLAME head vertices/faces are inserted. Faces in the transition band are rebuilt from a fixed triangulation so that the output template is watertight and has constant topology across frames. The default model applies neck-ring smoothing on FLAME's bottom line of 30 vertices. Mesh faces are manually predefined from the template assets; nearest-neighbor matching is used only for blend weights, not for triangulation.

Skinning weights. For vertices inherited from SMPL-X, we keep the original SMPL-X linear-blend-skinning weights. For FLAME face vertices, rigid head-region weights are assigned to the head joint except for jaw, eyeball, and transition-band vertices: FLAME jaw vertices keep jaw-dependent deformation, eyeball vertices follow the eye joints, and transition-band weights are interpolated from neighboring head and neck/torso weights. This preserves FLAME expression and jaw motion while making global head motion a child of the SMPL-X neck chain. FLAME jaw/eye skinning weights are copied directly from FLAME2023.

Forward pass. At inference, the network predicts $(\beta_h, \beta_t, \mathbf{s}, \psi, \boldsymbol{\theta}, \boldsymbol{\pi})$. PPM first builds the canonical injected template with FLAME shape/expression/jaw and SMPL-X torso shape, then applies the unified LBS under the spine $\rightarrow$ neck $\rightarrow$ head $\rightarrow$ jaw/eye hierarchy. Finally, vertices are projected with camera $\boldsymbol{\pi}$. Because the injection is performed in canonical space, the neck connection is independent of current head pose and remains stable under animation.

Computation and Manual Refinement of Barycentric Coordinates. To enable PPM to support projection of Sapiens2 facial and shoulder keypoints, we adopt a two-stage approach to compute the barycentric coordinates for each keypoint on the PPM mesh. First, we select a set of clear images as candidates. For each image, we use the Sapiens2 pose estimator to predict facial and upper body keypoints, and simultaneously use the data generation pipeline described in the next section (excluding losses related to Sapiens2 keypoints) to obtain a mesh that is aligned to the image. By projecting the estimated keypoints onto the mesh, we obtain coarse barycentric coordinates for each point.

Next, we develop a refinement tool that enables manual adjustment of these barycentric coordinates, ensuring that they precisely correspond to the Sapiens2 facial and shoulder keypoint positions and are consistent and accurate in 3D.

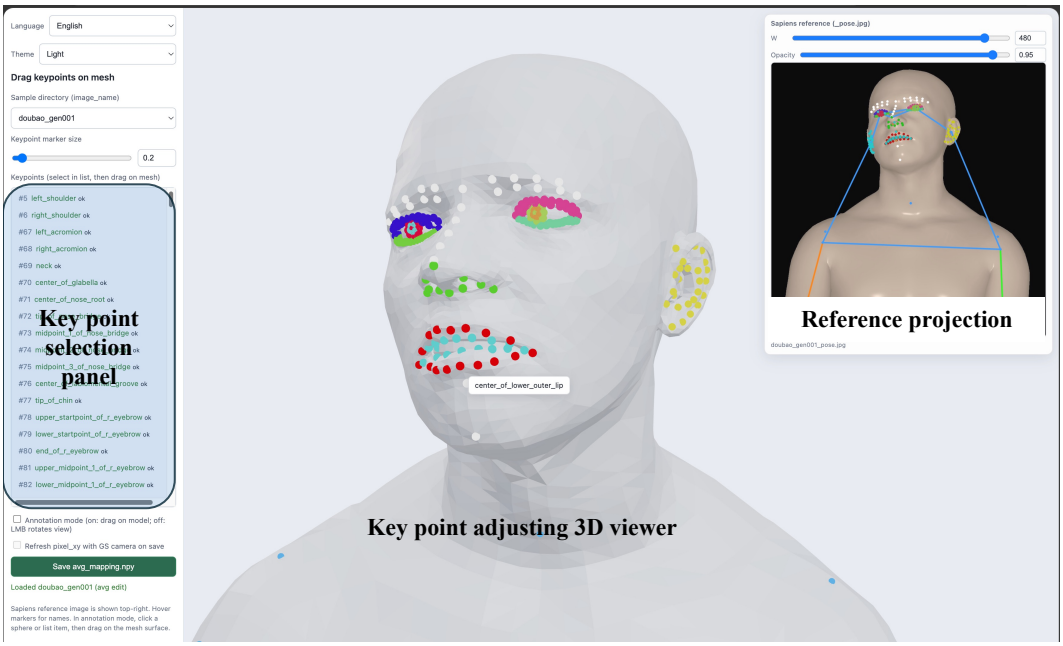


Figure 6: User interface for manually adjusting the barycentric coordinates of Sapiens2 keypoints on the PPM mesh. The tool allows refining the projected positions to ensure precise correspondence and 3D consistency.

C Data Generation Pipeline

Training PAA requires pseudo-labels that stage head pose, jaw motion, and expression. Most public datasets do not provide this, and joint fitting often mixes jaw rotation with expression blendshapes. We therefore run an offline **Graduated-Refinement Pipeline** on in-the-wild images (Fig. 7). The pipeline follows the same coarse-to-fine order as training: fit shape, then jaw, then expression, then full PPM parameters.

Each image $\mathbf{I}$ is processed in three phases: initialization, graduated FLAME fitting, and PPM integration.

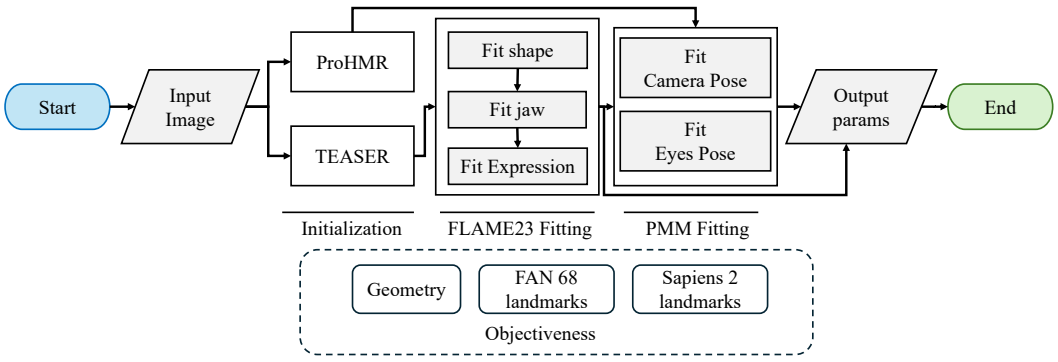


Figure 7: **Graduated-Refinement Pipeline.** We initialize body and face separately, then fit FLAME2023 in three steps (shape → jaw → expression), and finally integrate the result into PPM.

C.1 Phase 1: Hybrid Initialization

Fitting is sensitive to initialization. We warm-start from two regressors:

- **Body and camera (ProHMR):** ProHMR Kolotouros et al. (2021) gives initial camera $\boldsymbol{\pi}_{\text{init}}$ and torso shape $\boldsymbol{\beta}_{t,\text{init}}$.
- **Face (TEASER):** TEASER Liu et al. (2025) gives initial head shape $\boldsymbol{\beta}_{h,\text{init}}$, expression $\boldsymbol{\psi}_{\text{init}}$, and head-related pose.

C.2 Phase 2: Graduated FLAME-2023 Fitting

We fit FLAME2023 in three separate steps to reduce jaw–expression leakage (Fig. 7, center):

$$\mathbf{P}_{\text{face}} = \text{FLAME}(\boldsymbol{\beta}_h, \boldsymbol{\psi}, \boldsymbol{\theta}_{\text{jaw}}). \tag{10}$$

Step 2.1: Shape. Fix pose and expression to their initial values and optimize head shape only:

$$\min_{\boldsymbol{\beta}_h} \mathcal{L}_{\text{lmk}}(\text{FLAME}(\boldsymbol{\beta}_h, \boldsymbol{\psi}_{\text{init}}, \boldsymbol{\theta}_{\text{init}}), \mathbf{P}_{2d}) + \lambda_{\text{reg}} \|\boldsymbol{\beta}_h\|_2^2. \tag{11}$$

Step 2.2: Jaw. Fix $\boldsymbol{\beta}_h$ and set mouth-related expression coefficients to zero. Optimize jaw rotation $\boldsymbol{\theta}_{\text{jaw}}$ so mouth opening is explained by the jaw joint rather than expression.

Step 2.3: Expression. Fix $\boldsymbol{\beta}_h$ and $\boldsymbol{\theta}_{\text{jaw}}$, then optimize $\boldsymbol{\psi}$ for remaining surface detail (e.g., lip compression and cheek motion).

C.3 Phase 3: PPM Integration

We merge the fitted FLAME head into PPM and refine global pose.

Camera and torso. Using ProHMR body cues and the fitted head, we optimize $\boldsymbol{\pi}$, $\boldsymbol{\beta}_t$, and torso-related pose components in $\boldsymbol{\theta}$ (spine and neck). This aligns the spine → neck → head chain with the image.

Eyes. We refine eye gaze $\boldsymbol{\theta}_{\text{eye}}$ to match iris landmarks.

The final pseudo-label is

$$\mathbf{P}_{\text{out}} = \{\boldsymbol{\pi}, \boldsymbol{\beta}_h, \boldsymbol{\beta}_t, \boldsymbol{\theta}, \boldsymbol{\psi}\}, \tag{12}$$

which matches the parameterization in Section 3.1 and is used to supervise the PAA network during training.

D Training Objectives

This section expands the terms in Eq. (6). Mask-gated terms supervise the M- and L-experts only when their masks are on. $\mathcal{L}_{\text{lmk}}$ and $\mathcal{L}_{\text{fd}}$ stay active. Scalar weights are listed in Section B.

Graduated parameter loss ($\mathcal{L}_{\text{ord}}$). We apply L2 regression to pseudo-label parameters, with masks matching the Graduated-Mask Router:

$$\mathcal{L}_{\text{ord}} = \mathcal{L}_G(\boldsymbol{\pi}, \boldsymbol{\beta}_h, \boldsymbol{\beta}_t, \mathbf{s}) + m_M \mathcal{L}_M(\boldsymbol{\theta}_{\text{spine}}, \boldsymbol{\theta}_{\text{neck}}, \boldsymbol{\theta}_{\text{head}}, \boldsymbol{\theta}_{\text{jaw}}, \boldsymbol{\theta}_{\text{sh}}, \boldsymbol{\theta}_{\text{ua}}) + m_L \mathcal{L}_L(\boldsymbol{\psi}, \boldsymbol{\theta}_{\text{eye}}). \quad (13)$$

Each $\mathcal{L}_G$, $\mathcal{L}_M$, $\mathcal{L}_L$ is a standard L2 distance to the corresponding pseudo-label subset.

Feature distillation loss ($\mathcal{L}_{\text{fd}}$). We use channel-wise distillation to align student and teacher feature distributions rather than raw feature values. Let $\hat{\mathbf{Z}}, \mathbf{Z} \in \mathbb{R}^{N \times C}$ denote student and teacher token features (per sample), where N is the token dimension and C is the channel dimension. Following Giebenhain et al. (2025), we apply temperature-scaled softmax along the token dimension:

$$\mathbf{P}_s = \log \text{Softmax}\left(\frac{F(\hat{\mathbf{Z}})}{\tau}\right), \quad \mathbf{P}_t = \text{Softmax}\left(\frac{\mathbf{Z}}{\tau}\right), \quad (14)$$

and minimize the normalized KL divergence:

$$\mathcal{L}_{\text{fd}} = \frac{\tau^2}{NC} \text{KL}(\mathbf{P}_s \,\|\, \mathbf{P}_t). \quad (15)$$

Here $F(\cdot)$ is a 3-layer MLP projection head and τ is the distillation temperature. This objective transfers teacher attention over tokens and improves anatomy-aware token generalization for 3D portrait features.

Global 2D landmark loss ($\mathcal{L}_{\text{lmk}}$). Let $\mathbf{u}_j \in \mathbb{R}^2$ denote the j-th detected landmark and $\hat{\mathbf{u}}_j$ its projection from PPM. The G-Expert loss combines 68-point facial landmarks Zhu et al. (2016) (excluding the jawline, indices 0–16) and Sapiens-2 face/shoulder keypoints Khirodkar et al. (2026). We use the mean squared error:

$$\mathcal{L}_{\text{lmk}} = \frac{1}{|\mathcal{J}_G|} \sum_{j \in \mathcal{J}_G} \|\hat{\mathbf{u}}_j - \mathbf{u}_j\|_2^2, \quad (16)$$

where $\mathcal{J}_G$ indexes the supervised landmark set. This term anchors global face placement and head–shoulder layout.

Anatomical distance loss ($\mathcal{L}_{\text{nc}}$). Head pose is hard to supervise directly from sparse 2D points. To improve jaw–skeletal consistency, we use the nose-tip-to-chin-tip distance as a weak skeletal cue: under relatively stable head pose, it reflects jaw-driven chin motion more directly than lip-contour distances, which can be explained by local expression blendshapes. The targets are observed 2D landmarks. The same endpoints later define the JCR/JSR proxy, so JSR tracks this supervised cue.

$$\mathcal{L}_{\text{nc}} = |d(\hat{\mathbf{u}}_{\text{nose}}, \hat{\mathbf{u}}_{\text{chin}}) - d(\mathbf{u}_{\text{nose}}, \mathbf{u}_{\text{chin}})|, \quad (17)$$

where $d(\cdot, \cdot)$ is the Euclidean distance in the image plane, $\hat{\mathbf{u}}$ denotes the projected mesh landmarks, and $\mathbf{u}_{\text{nose}}, \mathbf{u}_{\text{chin}}$ are the observed 2D landmarks.

Overflow mask loss ($\mathcal{L}_{\text{om}}$). Torso vertices receive weaker 3D supervision than the face. We therefore penalize projected mesh regions that fall outside the portrait foreground mask $\mathcal{M}_{\text{fg}}$:

$$\mathcal{L}_{\text{om}} = \frac{|\Pi(\mathbf{V}_{\text{torso}}) \setminus \mathcal{M}_{\text{fg}}|}{|\mathcal{M}_{\text{fg}}|}, \quad (18)$$

where $\Pi(\cdot)$ is mesh rasterization in the image plane and $|\cdot|$ denotes pixel area. This keeps the body mesh inside the visible portrait region. Fig. 8 shows examples of the overflow mask.

Relative geometry loss ($\mathcal{L}_{\text{rg}}$). To add direct 3D supervision on the torso, we compare predicted and pseudo-label meshes in the camera frame after removing global translation. Let $\mathcal{V}_{\text{torso}}$ be the set of mesh vertices outside the injected FLAME head region (torso and shoulders). Denote by $\tilde{\mathbf{v}} = \mathbf{v} - \mathbf{t}$ the vertex position after subtracting the mesh translation. Then

$$\mathcal{L}_{\text{rg}} = \frac{1}{|\mathcal{V}_{\text{torso}}|} \sum_{\mathbf{v} \in \mathcal{V}_{\text{torso}}} \|\tilde{\mathbf{v}}_{\text{pred}} - \tilde{\mathbf{v}}_{\text{gt}}\|_2^2. \quad (19)$$

Excluding the FLAME head vertices avoids letting face fitting errors dominate the torso loss.

Table 5: Ablation on training objectives and Graduated-Mask Router. $\mathcal{L}_{\text{ord}}$ / $\mathcal{L}_{\text{fd}}$ / $\mathcal{L}_{\text{nc}}$ / $\mathcal{L}_{\text{om}}$ / $\mathcal{L}_{\text{rg}}$ follow the main loss terms; **Aug**: training-time augmentation; α_{act}: Bernoulli activation probability in the Graduated-Mask Router.

Variant	Aug	$\mathcal{L}_{ord}$	$\mathcal{L}_{fd}$	$\mathcal{L}_{nc}$	$\mathcal{L}_{om}$	$\mathcal{L}_{rg}$	α_{act}	NME-P ↓	NME-S ↓	HDTF JSR ↑	Ner-MVE ↓
V0.1 (ViT w/o DA)							1.0	0.141	0.559	0.270	37.1
V0.2 (ViT with DA)	✓						1.0	0.127	0.300	0.270	28.3
V0.3 (Sapiens2 Encoder)	✓						1.0	0.096	0.207	0.366	21.9
V1.1	✓	✓					1.0	0.062	0.176	0.676	21.0
V1.2	✓	✓	✓				1.0	0.064	0.172	0.683	14.4
V1.3	✓	✓	✓	✓			1.0	0.064	0.174	0.711	15.1
V1.4	✓	✓	✓	✓	✓		1.0	0.060	0.165	0.708	8.40
V1.5 (w/o router)	✓	✓	✓	✓	✓	✓	1.0	0.059	0.153	0.710	8.31
V2.1	✓	✓	✓	✓	✓	✓	0.75	0.058	0.156	0.784	8.19
V2.2 (Our final)	✓	✓	✓	✓	✓	✓	0.50	0.058	0.155	0.799	8.20
V2.3	✓	✓	✓	✓	✓	✓	0.25	0.066	0.165	0.801	8.58

Detail landmark losses. The L-Expert uses three landmark groups on fine facial regions:

$$\mathcal{L}_{\text{lmk}}^{\text{group}} = \frac{1}{|\mathcal{J}_{\text{group}}|} \sum_{j \in \mathcal{J}_{\text{group}}} \|\hat{\mathbf{u}}_j - \mathbf{u}_j\|_2^2, \quad \text{group} \in \{\text{mouth}, \text{eye}, \text{pupil}\}. \tag{20}$$

These terms supervise mouth shape, eye contours, and pupil locations.

Regularization. $\mathcal{L}_{\text{reg}}^{G}$, $\mathcal{L}_{\text{reg}}^{M}$, and $\mathcal{L}_{\text{reg}}^{L}$ are L2 penalties on the parameters predicted by each expert, which stabilizes regression when pseudo-labels are noisy.

E Additional Experimental Results

E.1 Full Component Ablation

We study training objectives and Graduated-Mask Router settings to answer **RQ4** (Table 5). Early rows (V0.x) replace the backbone or add data augmentation. Later rows add losses one by one, then change the router activation α_{act}. Among the reported JSR and Ner-MVE cells, V2.2 (Our final, α_{act}=0.50) reaches JSR 0.799 and Ner-MVE 8.20, which matches the full model in the main tables. V1.5 (α_{act}=1) is the same run as "w/o Graduated-Mask Router" in Table 4 (JSR 0.710, Ner-MVE 8.31).

Fig. 8 shows the overflow mask used in $\mathcal{L}_{\text{om}}$. It penalizes projected torso pixels outside the foreground mask and helps keep the body mesh inside the visible portrait region.

E.2 Jaw–Skeletal Ratio Analysis

Fig. 5 (main paper) visualizes jaw–skeletal coupling on a representative talking clip. As annotated in that figure, the *jaw-chain ratio* (JCR) on frame t is d_2/d_1. d_1 is the horizontal face width, measured between the left and right facial boundaries. d_2 is the nose-tip-to-chin distance (68-point indices 33 and 8):

$$d_1^t = \|\mathbf{u}_{\text{face,l}}^t - \mathbf{u}_{\text{face,r}}^t\|_2, \quad d_2^t = \|\mathbf{u}_{\text{nose}}^t - \mathbf{u}_{\text{chin}}^t\|_2, \quad \text{JCR}_t = \frac{d_2^t}{d_1^t}. \tag{21}$$

Dividing by face width removes subject scale and crop size while keeping the skeletal opening trend from jaw motion. We define the *jaw–skeletal ratio score* (JSR) as

$$\text{JSR} = \rho\big(\{\theta_{j-p}^t\}_{t=1}^T, \{\text{JCR}_t\}_{t=1}^T\big), \tag{22}$$

where θ_{j-p}^t is predicted jaw pitch and $\rho(\cdot,\cdot)$ is Pearson correlation. JSR is the skeletal-consistency proxy used for RQ2. $\mathcal{L}_{\text{nc}}$ supervises the unnormalized distance d_2 against observed 2D landmarks, and JCR uses the same nose and chin endpoints, normalized by d_1. A higher JSR means predicted jaw pitch tracks this observed cue. The main paper reports HDTF-wide JSR in Table 2.

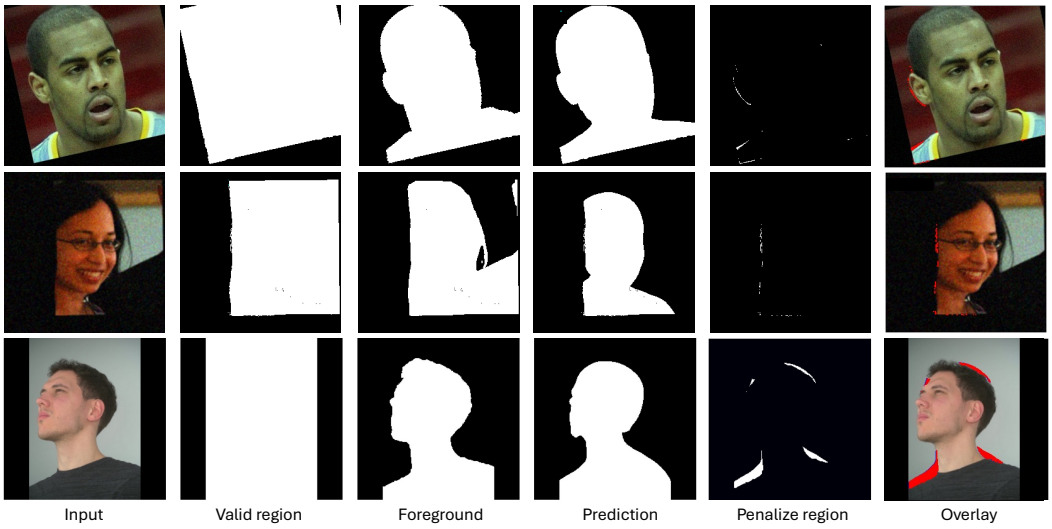


Figure 8: Visualization of the overflow mask used by $\mathcal{L}_{\text{om}}$. The mask marks projected torso regions that fall outside the portrait foreground.

E.3 Additional Jaw–Expression Qualitative Results

Fig. 9 shows a high mouth-opening clip complementary to Fig. 5. Compared with TEASER Liu et al. (2025), PEAR Wu et al. (2026), and SPECTRE Filntisis et al. (2023), GRAPE assigns large mouth opening more consistently to jaw articulation rather than to expression shortcuts. The **deep-red box** marks a case that is anatomically inconsistent: the jaw should close the mouth while expression opens it, yet both open the mouth. The **deep-blue box** marks a case where the jaw should drive opening, but the predicted jaw under-opens and expression dominates. The same colors mark extreme Var-Exp / Var-Jaw entries in Table 2.

E.4 Expert Attention

Fig. 10 visualizes expert attention under the graduated parameter loss $\mathcal{L}_{\text{ord}}$. With this loss, different experts attend to the image regions that match their anatomical roles.

E.5 Metric Definitions and Evaluation

Metric Definitions. The metrics are defined as follows:

- **68pt NME** (↓): inter-ocular normalized mean error in %, excluding jawline indices 0–16 Wu et al. (2026).
- **Sapiens-Face / Shoulder** (↓): mean L2 in normalized crop coordinates on Sapiens keypoint subsets Khirodkar et al. (2026).
- **JCR**: per-frame ratio d_2/d_1, where d_1 is the face width and d_2 is the nose-tip-to-chin distance (Fig. 5). It is a scale-normalized skeletal mouth-opening proxy on clips with relatively stable head pose.
- **JSR** (↑): Pearson r between predicted jaw pitch and JCR. This is the skeletal-consistency proxy for RQ2. JCR reuses the observed nose–chin landmarks supervised by $\mathcal{L}_{\text{nc}}$ (Fig. 5, Section E.2).
- **Temporal** (↓): mean second-order temporal difference of jaw and expression parameters.
- **NME-P / NME-S** (↓): projection landmark errors on face / shoulder regions in the talking-head eval crop.
- **MVE / LVE** (↓): mean / lip vertex error on the shared FLAME2020/FLAME2023 head topology Feng et al. (2021b). Neck and torso vertices are excluded.



Figure 9: **Jaw versus expression on a talking clip.** The **deep-red box** highlights a case that is anatomically inconsistent: the jaw should close the mouth while expression opens it, yet both jaw and expression open the mouth. The **deep-blue box** highlights a case where the jaw should drive mouth opening, but the predicted jaw under-opens and expression dominates.

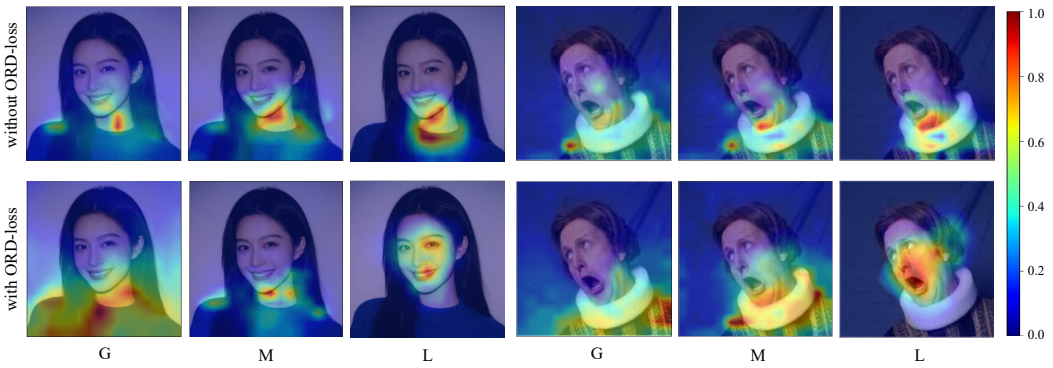


Figure 10: **Expert-net attention visualization.** With $\mathcal{L}_{\text{ord}}$, our expert nets focus on the accurate regions of the input image.

Table 6: Quantitative avatar reconstruction with RGBAvatar Li et al. (2025) on HDTF. We compare avatars driven by FLAME2020 parameters from TEASER Liu et al. (2025), GRAPE head-only parameters (`ppm_head`), and full upper-body PPM parameters (`ppm_ub`). **SSIM / PSNR**: full-frame reconstruction quality (↑). **SSIM-b / PSNR-b**: body-region reconstruction quality (↑).

Method	SSIM ↑	PSNR ↑	SSIM-b ↑	PSNR-b ↑
FLAME2020 (TEASER Liu et al. (2025))	0.844	21.08	0.673	17.90
PPM head-only (GRAPE)	<u>0.845</u>	<u>21.39</u>	<u>0.687</u>	<u>18.23</u>
PPM (GRAPE)	**0.848**	**21.69**	**0.693**	**18.64**

- **MPJPE** (↓): joint error in millimeters over the head and both shoulders Kolotouros et al. (2019); Liu et al. (2025). Shoulder joints make the absolute value larger than face-only NME.

Evaluation Protocol. All image and video baselines use their own official crop protocols for fair comparison (Section 4.1). HDTF Zhang et al. (2021) evaluation uses 20 randomly selected videos, consistent with the main paper. FLAME2020 and FLAME2023 share the same vertex topology, so Nersemble MVE/LVE compare methods on that common mesh. The Nersemble test set does not overlap the training set.

E.6 More Downstream Results

Audio-driven talking-head generation. Following DiffPoseTalk Sun et al. (2024), we build two training sets from the same 10 HDTF clips: one from TEASER Liu et al. (2025) parameters and one from GRAPE parameters. Fig. 11 shows the main training-loss comparison. The GRAPE-labeled set shows a modest training improvement. Together with the RGBAvatar reconstruction numbers, this reflects improved training/reconstruction behavior of the downstream models.

Animatable 3D Gaussian avatar generation. We train RGBAvatar Li et al. (2025) on three settings from the same HDTF sequences: (1) TEASER-generated parameters, (2) GRAPE head-only parameters (same head mesh topology as TEASER), and (3) full GRAPE parameters. Fig. 12 compares the reconstructed avatars against the reference. The TEASER-based avatar shows clearer artifacts around the shoulders due to the floating-head issue, while full GRAPE keeps a more continuous head–neck–shoulder surface.

We further report quantitative RGBAvatar results in Table 6. We randomly select 10 videos; for each video, the last 350 frames are used as the test set and the remaining frames as the training set. Full PPM slightly improves full-frame SSIM/PSNR over TEASER. Body-region PSNR is 18.64, against 18.23 for head-only PPM and 17.90 for TEASER, alongside the neck–shoulder continuity in Fig. 12.

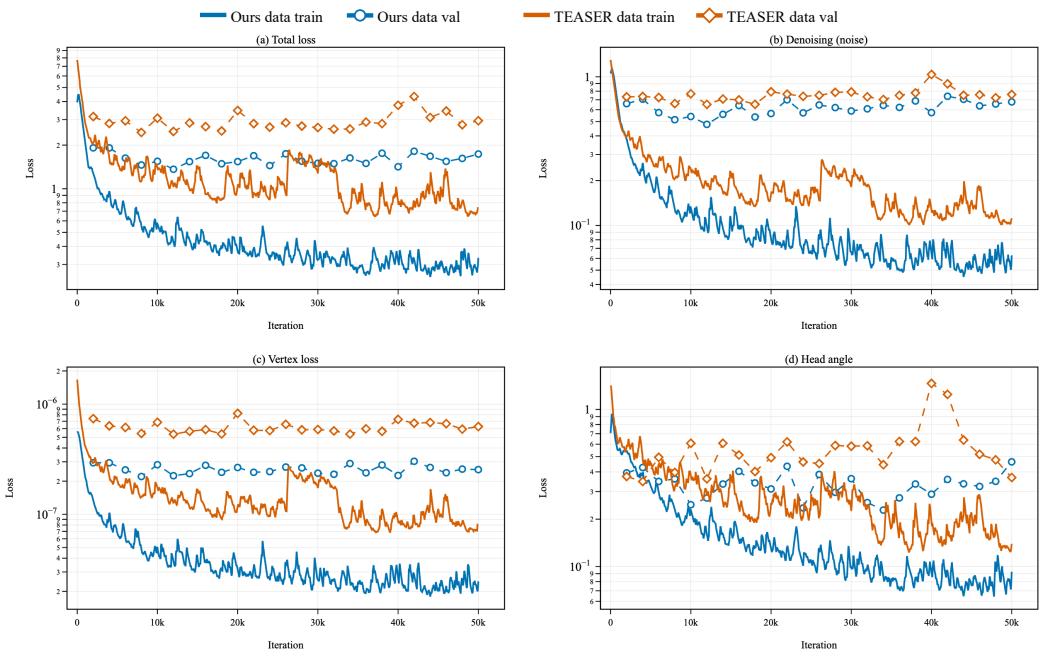


Figure 11: **Talking-head training on HDTF.** We train DiffPoseTalk Sun et al. (2024) on TEASER-generated parameters and on GRAPE-generated parameters. The loss curves show a modest training improvement with the GRAPE-labeled set.

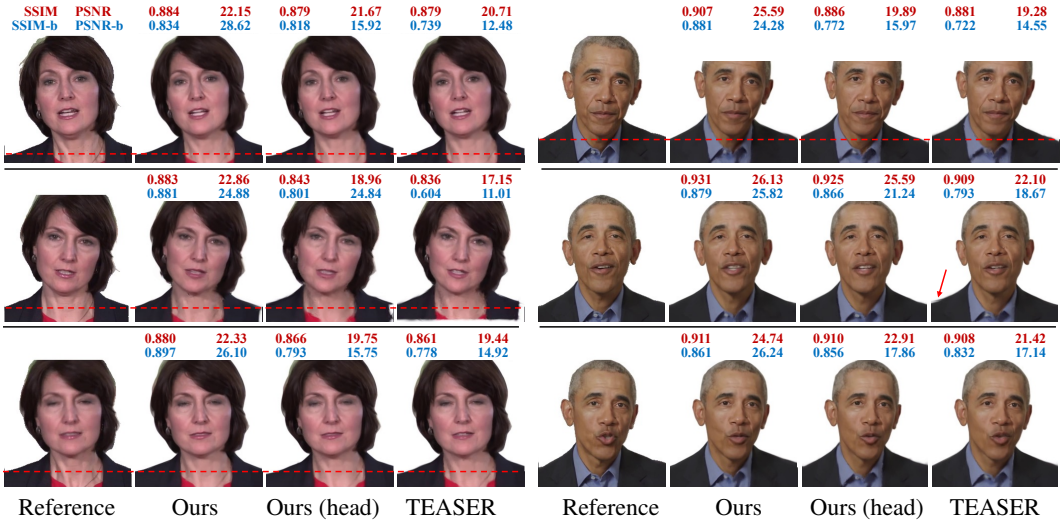


Figure 12: **Avatar generation and animation on HDTF.** Based on RGBAvatar Li et al. (2025), we compare avatars trained on TEASER parameters, GRAPE head-only parameters, and full GRAPE parameters. Relative to TEASER, full GRAPE reduces the "floating head" artifact and keeps better continuity in the neck–shoulder region. Here SSIM-b and PSNR-b indicate the SSIM and PSNR of the body part only.

F LIMITATIONS, DISCUSSIONS AND FUTURE WORK

Limitations. GRAPE targets portrait images where the face, neck, and at least part of the shoulders are visible. The torso-rooted PPM improves head–neck coherence when these regions are observed. The benefit can drop when the neck or shoulder is heavily occluded by hair, hands, scarves, or loose clothing. PPM is a tight-surface parametric model. Constrained by the FLAME representation, the current PPM does not accurately model teeth, the tongue, or other intra-oral structures, and thus cannot fully capture the full range of portrait expressions. It does not reconstruct hair volume, garments, jewelry, or other non-body geometry. In those cases, the projected mesh can still serve as a kinematic scaffold, but it is not a full scene reconstruction.

Training and evaluation also depend on pseudo supervision. Sapiens-2 keypoints, 68-point landmarks, foreground masks, and offline-fitted PPM labels provide complementary signals. Each signal can fail under extreme profile views, motion blur, low resolution, unusual lighting, or uncommon poses. Multi-source supervision reduces reliance on any single cue, but it cannot remove all pseudo-label bias. JCR/JSR is a skeletal-consistency proxy for talking clips with relatively stable head pose. It reuses the observed nose–chin landmarks supervised by $\mathcal{L}_{\mathrm{nc}}$, and we read it together with qualitative shortcutting checks. Large head motion, detector jitter, or bad nose/chin localization can affect the score. JSR should be read together with qualitative results and reconstruction metrics, not alone.

Discussions. The main design choice is to treat portrait recovery as an articulated portrait problem, not a face-only fitting problem. PPM closes the representation gap with a torso-rooted reference frame. Multi-source supervision addresses weak observability of portrait variables. PAA with the Graduated-Mask Router is the training formulation that keeps jaw pitch consistent with skeletal mouth opening and reduces expression shortcuts. These parts work together. A richer model without stronger supervision can still drift in the shoulder region. Stronger supervision without anatomical routing can still fit mouth motion through expression shortcuts. The graduated design is most useful when the output will be animated or edited, where parameter meaning matters beyond per-frame landmark accuracy.

There is also a trade-off between generality and controllability. Implicit or generative representations can capture hair, clothing, and fine appearance more easily, but they often give weaker explicit control over neck, head, and jaw. PPM keeps an explicit kinematic structure for talking-head animation and avatar control, at the cost of modeling only the body-like portrait surface.

Future Work. We already observe improved training/reconstruction behavior for DiffPoseTalk and RGBAvatar in the main experiments (Section 4.5); RGBAvatar quantitative metrics are in Table 6. Future work can extend GRAPE in three directions. First, combine the PPM scaffold with neural implicit fields or Gaussian splatting to model hair, clothing, and accessories while keeping kinematic control. Second, improve pseudo-label generation with temporal fitting, multi-view consistency, or uncertainty-aware supervision, so unreliable shoulder and chin cues weigh less during training. Third, test jaw–skeletal consistency beyond the JCR/JSR proxy, using tracked jaw joints, dental scans, or high-quality 4D face captures.