

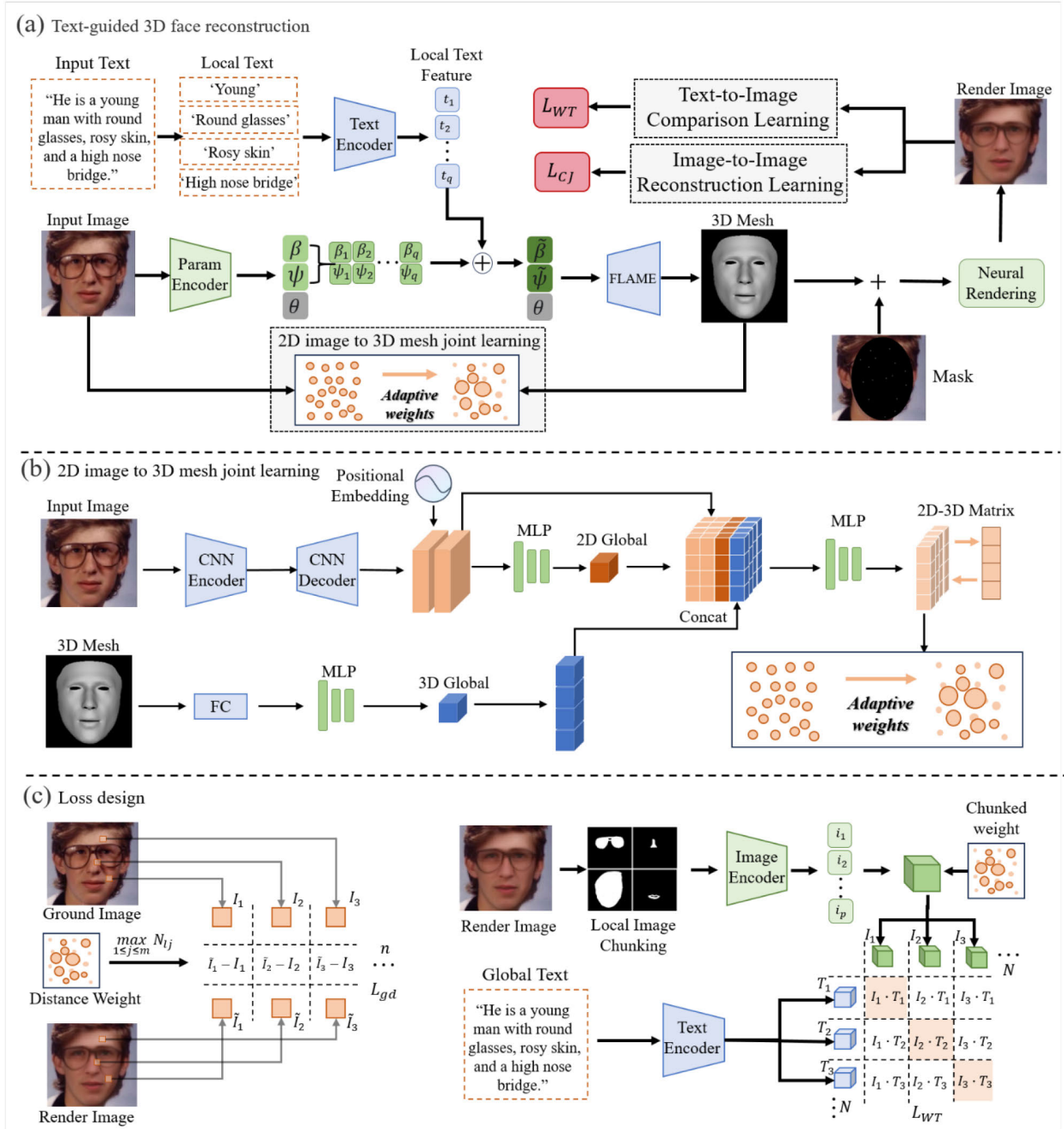


Fig. S1 Overview of the AsmV2L framework. (a) Text-Guided 3D Face Reconstruction, where structured facial descriptions are dynamically generated from images of faces and then used as semantic prompts to modulate the parameters of the 3D model, guiding the 3D face reconstruction. **(b) Joint Learning of 2D Images and 3D Meshes**, which integrates 2D image features with 3D mesh features and establishes pixel-to-vertex spatial mapping via a shared multilayer perceptron. **(c) Training Loss**, where an adaptive multi-scale loss is designed to achieve high-fidelity reconstruction.

Materials and methods

We propose AsmV2L, a 3D face reconstruction method guided by fine-grained text semantics and 2D-to-3D spatial mapping, leveraging vision-language models. An illustrative overview of the framework is presented in Fig. S1a. We first constructed structured text descriptions, then guided the parametric model to reconstruct 3D faces based on the correspondence between text sub-features and facial regions. Finally, by jointly learning 2D images and 3D meshes, we established pixel and vertex-level mapping relationships to adaptively adjust multi-scale loss functions, achieving high-precision 3D face reconstruction under semantic constraints.

1 Text-guided 3D face reconstruction

1.1 Extracting text descriptions from face images

Given an input image $I \in R^{H \times W \times 3}$, we utilized the bootstrapping language-image pre-training (BLIP) model (Chen et al., 2025) to generate textual information describing the facial image across multiple dimensions: basic information J (gender, skin tone, age), shape information X (nose, mouth, eyes, face shape, expression), and texture information L (accessories, beard). A text template T was then constructed from these: “A [skin tone] [age] [gender] with facial features including [nose], [mouth shape], [eyes], and [face shape]; facial textures include [accessories] and [beard], conveying a [expression] expression.” Subsequently, embeddings of the text and image were extracted using the CLIP text encoder E_T and CLIP image encoder E_I (Alec et al. 2021), respectively.

1.2 Face model

We employed the Faces Learned with an Articulated Model and Expressions (FLAME) model (Li et al., 2017) to generate 3D faces from a single 2D facial image. FLAME is a parametric model that controls rotations around four key joints (left eyeball, right eyeball, chin, neck) through parameters, including shape parameter β , pose parameter θ , and expression parameter ψ . A face parameter encoder is used to regress FLAME parameters: $\text{ParamEncoder}(I) \rightarrow \beta \in R^{|\beta|}$, $\theta \in R^{|\theta|}$, $\psi \in R^{|\psi|}$, and $c \in R^3$, where c denotes camera parameters that determine the projection direction when rendering images from 3D meshes.

1.3 Reconstructing 3D faces from images

To enhance the precision of text-guided parameter modulation, we implemented a fine-grained text-guided 3D face reconstruction strategy, which adaptively assigns weights to different facial regions based on textual semantics, targeting optimization of key regions emphasized in text prompts (e.g., “upturned mouth corners”) and addressing the limitations of global parameter fusion in capturing fine-grained semantic alignment.

The original shape parameter β and expression parameter ψ were divided into q semantically meaningful chunks, each corresponding to a specific facial region. Formally, these are defined as:

$$\beta = \{\beta_1, \beta_2 \dots \beta_q\}, \psi = \{\psi_1, \psi_2 \dots \psi_q\}, \beta_i, \psi_i \in R^{\frac{d}{q}}, \quad (1)$$

where β_i and ψ_i represent subsets of parameters for the i th facial region. Fine-grained textual attributes are processed by the text encoder E_T to extract textual feature $\tilde{T} = \{t_i\}_{i=1}^q \in R^{512 \times q}$, which is further projected into a matrix $Z = L_Z(\tilde{T}) \in R^{512 \times q}$ via learning a linear projection L_Z .

To align text with shape/expression parameters, two independent mapping functions Y_i^β and Y_i^ψ are constructed to compute attention weights, measuring the correlation between text semantics and visual parameters. The chunk weights for the shape and expression parameters are denoted as ω_i^β and ω_i^ψ , respectively:

$$\omega_i^\beta = \text{softmax}\left(\frac{Y_i^\beta(Z) \cdot \beta_i^T}{\sqrt{d/q}}\right) \quad (2)$$

$$\omega_i^\psi = \text{softmax}\left(\frac{Y_i^\psi(Z) \cdot \psi_i^T}{\sqrt{d/q}}\right), \quad (3)$$

where $\sqrt{d/q}$ serves as a scaling factor. The final modulated parameters $\bar{\beta}$ and $\bar{\psi}$ are obtained by weighted concatenation of chunk parameters:

$$\bar{\beta} = \text{concat}(\omega_i^\beta \cdot \beta_i)_{i=1}^q, \bar{\psi} = \text{concat}(\omega_i^\psi \cdot \psi_i)_{i=1}^q, \quad (4)$$

The modulated parameters $\bar{\beta}$ and $\bar{\psi}$ are fed into the FLAME model to generate a 3D face mesh M :

$$\text{FLAME}(\beta, \theta, \psi) \rightarrow M \in (V, F), \quad (5)$$

where $V \in R^{m_v \times 3}$ represents 3D mesh vertices $m_v = 5023$ and $F \in R^{m_f \times 3}$ represents 3D mesh faces $m_f = 9976$.

2 Joint learning of 2D images and 3D meshes

The framework for joint learning between 2D images and 3D meshes is outlined in Fig. S1b. For 2D image features, a fully convolutional encoder E_c and decoder E_d are employed to extract image feature f_{tx} from the input facial image. To enhance geometric information in 2D features, a 2D position encoder E_p is used to extract 2D position feature f_{wz} . These 2D image features and position features are fused to form 2D local feature: $f_{local} = f_{tx} \oplus f_{wz}$. The 2D global feature is derived by applying global average pooling to the 2D local feature: $\text{GAP}(f_{local}) \rightarrow f_{global}$. Finally, the local and global features are concatenated to form the complete 2D feature f_{2D} :

$$f_{local} \oplus f_{global} \rightarrow f_{2D}. \quad (6)$$

Each vertex in the 3D mesh $V \in R^{m_v \times 3}$ is represented by its 3D spatial coordinates, which contain rich geometric information, and 3D feature f_{3D} is extracted from the 3D mesh V as follows:

$$\text{GAP}\left(\text{MLP}\left(\text{FC}(V)\right)\right) \rightarrow f_{3D}, \quad (7)$$

where $\text{FC}(\cdot)$ denotes a fully connected layer, $\text{MLP}(\cdot)$ denotes a multi-layer perceptron, and $\text{GAP}(\cdot)$ represents global average pooling.

The 2D feature f_{2D} and 3D feature f_{3D} are concatenated and input into a multilayer perceptron (MLP) to jointly learn the probability distribution $N_{i,j} \in R^{n \times m_v}$ mapping 2D pixels to 3D spatial vertices. Here, n is the number of pixels, and m_v is the number of 3D vertices:

$$\text{softmax}\left(\text{MLP}(f_{2D}, f_{3D})\right) \rightarrow N_{i,j}, \quad (8)$$

where $\text{softmax}(\cdot)$ is used for normalization. In the probability matrix $N_{i,j}$, i denotes a pixel and j denotes the probability that each pixel corresponds to a 3D mesh vertex.

3 Training loss

3.1 Rendering reconstructed facial images

We constrain the 3D mesh by reconstructing facial images. For shape information, differentiable rendering $W(\cdot)$ is used, which outputs a monochromatic rendered image D given the 3D face mesh and camera parameter c : $W(M, c) \rightarrow D$.

For textural information, instead of traditional UV texture maps, neural rendering $S(\cdot)$ is employed to learn textural details from randomly sampled pixels. To avoid over-reliance on sampled pixels, the facial region is segmented from the original image, and a sparse sampling strategy $M(\cdot)$ is applied to sample pixels $M(I)$ from this region. Finally, neural rendering outputs the reconstructed facial image:

$$S(M(I) \oplus D) \rightarrow \tilde{I}. \quad (9)$$

3.2 Loss between real and reconstructed images

Photometric loss constrains the 3D mesh by measuring pixel-wise errors between the input image I and the rendered image $\tilde{I}$. In 3D face reconstruction, pixels closer to the projection of 3D vertices play a more critical role. Based on this, we calibrate pixel confidence to differentiate their contributions. The maximum probability of each pixel in the probability matrix, $\max_{1 \leq j \leq m} N_{ij}$, is used as confidence to weight the photometric loss, achieving fine-grained pixel-level constraints:

$$L_{gd} = \frac{1}{n} \sum_{i=1}^n \max_{1 \leq j \leq m_v} N_{ij} \|V_i \odot (\tilde{I} - I)\|_1, \quad (10)$$

where V_i denotes the rendering mask of the output face shape (1 for facial region pixels, 0 otherwise), and $\odot$ is the Hadamard product.

We adopt landmark consistency loss L_{db} to minimize the Euclidean distance between real and reconstructed 2D facial landmarks. Expression consistency loss L_{qg} uses a pre-trained emotion network to quantify feature differences between input and rendered images. L_ψ is the expression parameter regularization. The final image reconstruction loss is:

$$L_{CJ} = \lambda_{gd} L_{gd} + \lambda_{db} L_{db} + \lambda_{qg} L_{qg} + \lambda_\psi L_\psi. \quad (11)$$

3.3 Loss between text and reconstructed images

To enhance semantic consistency between the generated 3D face and text descriptions, we focused on aligning local attributes between text and reconstructed images, as shown in Fig. S1c. Facial attributes are primarily manifested in specific image regions, with different regions making different contributions. We therefore calibrated the confidence of different image regions to emphasize fine-grained semantic roles. Specifically, based on fine-grained text chunks, pixels belonging to the same facial region in the probability matrix $N_{i,j}$ are aggregated, and the weighted average yields the weight for that region: $\{\omega_k\}_{k=1}^q = \frac{1}{n/q} \sum_{i=1}^{n/q} \max_{1 \leq j \leq m} N_{ij}$. The text-to-image loss $L(x_i)$ is defined using these regional weights:

$$L(x_i) = -\frac{1}{N} \log \frac{\exp\left(\sum_{k=1}^q \omega_k \cdot E_T(s_i) \cdot \frac{E_I^k(x_i)}{\tau}\right)}{\sum_{j=1}^N \exp\left(\sum_{k=1}^q \omega_k \cdot E_T(s_j) \cdot \frac{E_I^k(x_i)}{\tau}\right)}, \quad (12)$$

where E_T and E_I are text and image encoders, respectively; τ is a learnable parameter; N is the batch size; s_i is the positive image sample generated from the input text; and the other generated images serve as negative samples s_j . Similarly, positive and negative texts can be identified for image x_i , yielding the image-to-text loss $L(s_i)$. The final contrastive loss is:

$$L_{WT} = \frac{1}{2N} \sum_{i=1}^N [L(s_i) + L(x_i)]. \quad (13)$$

4 Disease representation based on 3D facial expression dynamics

Given an input facial video $X = \{x_t\}_{t=1}^T$, AsmV2L first reconstructs frame-wise expression-related 3DMM coefficients, which are used as frame-level affect descriptors because they better capture non-rigid facial changes while reducing the influence of identity, illumination, and other appearance variations. Let ψ_t denote the expression coefficient vector reconstructed by AsmV2L at frame t . The coefficient sequence $\{\psi_t\}_{t=1}^T$ is fed into a temporal regression head to estimate frame-level valence and arousal:

$$e^t = G(\{\psi_t\}_{t=1}^T)_t = [v^t, a^t], \quad (14)$$

where $G(\cdot)$ denotes a pre-trained temporal regressor comprising an LSTM encoder followed by a four-layer MLP, and e^t denotes the predicted valence-arousal pair of the t th frame. To support downstream affective analysis, we further computed dynamic statistics from the reconstructed coefficient sequence and the VA sequence, and concatenated them to form a compact segment descriptor:

$$h = \text{Stat}(\{\psi_t\}_{t=1}^T) \parallel \text{Stat}(\{e^t\}_{t=1}^T), \quad (15)$$

where $\text{Stat}(\cdot)$ denotes temporal statistical encoding and $\parallel$ denotes feature concatenation. For patient-level analysis, segment descriptors from the same subject are further aggregated to obtain a patient representation for BD–MDD discrimination.

References

- Chen J, Xu Z, Pan X, et al., 2025. BLIP3-o: a family of fully open unified multimodal models-architecture, training and dataset. arXiv:2505.09568.
<https://doi.org/10.48550/arXiv.2505.09568>
- Li T, Bolkart T, Michael J, et al., 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)*, 36(6):194.
<https://doi.org/10.1145/3130800.3130813>
- Alec R, Jong H, Chris H, et al., 2021. Learning transferable visual models from natural language supervision. arXiv:2103.00020.
<https://doi.org/10.48550/arXiv.2103.00020>