GEOMETRY-CONDITIONED ADAPTATION FOR CLIP-BASED CROSS-DOMAIN FACIAL BEAUTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial beauty prediction (FBP) estimates ratings from particular annotator populations, yet imaging conditions and rating protocols vary across datasets. Landmark geometry encodes facial proportions and anatomical relationships without pixel-level texture or illumination, offering a complementary structural cue that may be less sensitive to changes in image appearance. This could improve robustness to domain shifts across datasets. Source-tuned CLIP provides transferable, language-aligned representations, but directly fusing image-derived landmarks may amplify noisy or redundant features and perturb the image–text alignment used for prompt-based scoring. We propose a geometry-conditioned tangent adapter that keeps the CLIP image embedding as the prediction anchor. A separately trained landmark–contour encoder supplies structural features that condition image-dependent corrections. With both encoders frozen, the adapter projects its correction onto the tangent plane, bounds its norm, and re-normalizes the image embedding before applying the original text prompts and scoring rule. Source-training statistics calibrate adapter inputs, and zero initialization recovers the frozen scorer before adapter training. We evaluate on LiveBeauty, MEBeauty, and SCUT-FBP5500 under three within-dataset and two cross-dataset protocols without target-domain adaptation. The full model improves all three correlation metrics over the reported state-of-the-art FPEM results across all five settings. For LiveBeauty-to-SCUT-FBP5500 transfer, the full system's relative SRCC, PLCC, and KRCC gains over FPEM are 23.5%, 32.9%, and 26.0%, respectively. In the same setting, PLCC rises from 0.726 for our source-tuned CLIP base to 0.767 for the full model ($+5.65\%$ relative to that base). These results support bounded geometry-conditioned adaptation as a useful design for this CLIP-based FBP system without target-domain fine-tuning.

Keywords: facial beauty prediction; vision–language models; facial landmark graphs; geometry-conditioned adaptation.

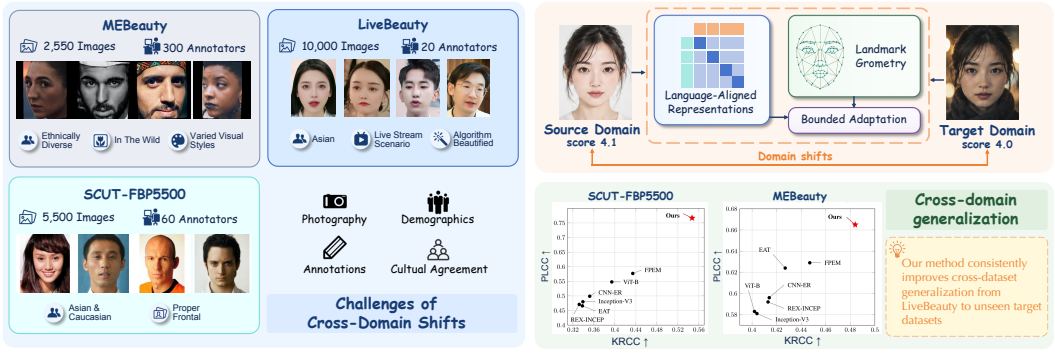

Figure 1: Overview of the cross-domain facial beauty prediction challenge. Our method uses facial landmark geometry to condition a bounded update of the language-aligned CLIP representations, improving generalization to unseen target datasets.

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.

1 INTRODUCTION

Facial beauty prediction (FBP) estimates aggregate judgments from the annotators and face populations represented in a dataset. Although these judgments exhibit substantial agreement, they vary with face and observer populations, while dataset labels depend on rating scales and annotation protocols (Langlois et al., 2000; Coetzee et al., 2014; Li et al., 2025; Lebedeva et al., 2022; Liang et al., 2018). Cross-dataset evaluation asks whether a scorer trained under one such combination transfers to another.

Vision–language models (VLMs) offer a promising alternative to direct score regression by aligning facial appearance with semantic attractiveness descriptions. CLIP learns transferable image–text representations (Radford et al., 2021), while FPEM adapts CLIP with face priors and ordered language prompts for FBP (Li et al., 2025). Yet strong within-dataset FBP performance does not ensure cross-dataset transfer: LiveBeauty contains potentially retouched live-streaming faces, MEBeauty includes ethnically diverse in-the-wild faces rated on a different scale, and SCUT-FBP5500 contains Asian and Caucasian frontal faces (Li et al., 2025; Lebedeva et al., 2022; Liang et al., 2018). These differences create shifts in demographics, photographic conditions, image processing, annotator populations, and score distributions. A source-trained scorer may consequently rely on texture, makeup, lighting, or processing cues that do not persist across datasets (Zhou et al., 2023; Geirhos et al., 2020).

Facial landmarks are widely used in face analysis as sparse descriptions of facial shape and spatial relationships (Zhang et al., 2011; Liu et al., 2017a; Zhao et al., 2024). Their relative positions describe proportions and symmetry relevant to attractiveness without directly encoding pixel texture or illumination. These structural cues may remain useful when appearance cues learned by a source-tuned CLIP scorer change across datasets, motivating geometry as a complementary input for cross-dataset prediction. Landmarks are nevertheless estimated from images and can inherit detection errors or domain biases; they also cannot resolve changes in subjective rating criteria. CLIP already captures some facial structure, but its global language-aligned embedding has no direct correspondence with node-structured landmark features. Fusing these representations without accounting for their different structure and reliability may introduce noisy or redundant geometry and degrade prediction. The challenge is to let facial topology complement CLIP while limiting its effect on the established image representation.

To address these challenges, we retain the source-tuned CLIP image embedding as the prediction anchor and train a dual-stream graph encoder on source data to extract landmark and contour features; source-centered node tokens also form regional descriptors. We then freeze the CLIP and geometry encoders and calibrate the adapter inputs using source-training statistics. An image-only branch is combined with three low-rank geometry branches, each of which multiplicatively couples image context with landmark, contour, or region features to produce a sample-dependent correction rather than a separate geometry score. We project the combined correction onto the CLIP embedding's tangent plane, bound its norm, and re-normalize the updated embedding, thereby limiting its angular displacement. The original text prompts, logit scale, and scoring rule then evaluate the adapted embedding; zero initialization makes the adapter start from the frozen CLIP scorer.

Our main contributions are summarized as follows:

- We formulate cross-domain FBP adaptation around a source-tuned CLIP scorer, with facial topology serving as conditional structural information.
- We introduce a geometry-conditioned tangent adapter that combines an image-only correction with multiplicative image–geometry interactions, then projects and bounds the resulting update in the CLIP embedding's tangent space.
- Our method exceeds FPEM on all three correlation metrics across three within-dataset and two cross-dataset protocols (Li et al., 2025). On LiveBeauty-to-SCUT-FBP5500 transfer, its relative SRCC, PLCC, and KRCC gains over FPEM are 23.5%, 32.9%, and 26.0%, respectively. Against our frozen source-tuned CLIP base in the same setting, PLCC rises from 0.726 to 0.767 ($+5.65\%$).

2 Related Work

2.1 Facial Beauty Prediction

Early FBP used facial proportions, symmetry, and appearance descriptors; later models learned visual features and rating distributions (Eisenthal et al., 2006; Zhang et al., 2011; Schmid et al., 2008; Fan et al., 2018; Lin et al., 2019). Geometric measurements summarized facial shape, while image models also captured texture and other appearance cues (Liu et al., 2017a; Gray et al., 2010; Bougourzi et al., 2022). Their targets are ratings from particular annotator groups and protocols, not universal attractiveness values.

Vision–language approaches align faces with attractiveness terms: ITM-MGFA learns multi-granularity image–text alignment, while FPEM combines face priors and semantic features through selective fusion and an ordered-prompt scorer (Chen et al., 2025; Li et al., 2025). Feature selection, fusion, and prompt-based scoring are therefore established in FBP. Our source-tuned CLIP base follows the prompt readout of FPEM's Multi-modal Attractiveness Encoder (MAEM), but is trained in our pipeline and differs from the published full FPEM model. We condition a bounded update of that base image embedding on dense geometry while retaining its prompts, logit scale, and scoring rule.

2.2 Graph-based Facial Landmark Representation

Landmark trajectories support temporal facial analysis (Jung et al., 2015); graph-based expression methods separately encode anatomical and temporal relations among facial points (Zhang et al., 2017; Zhou et al., 2020; Jin et al., 2021; Xin et al., 2021). Specified or learned graph edges make local facial relationships explicit, although these studies mainly address expression and motion. LBAGCN models landmark and virtual-muscle graphs, while FineCLIPER adds landmark maps as a CLIP modality (Zhao et al., 2024; Chen et al., 2024). For single-frame FBP, we learn landmark, contour, and anatomical-region representations from source labels, then freeze the geometry encoder and use its features to condition an ordinal image–text scorer. The prompt-based scoring path itself remains fixed during this adaptation stage.

2.3 Geometry–Appearance Fusion and Feature Adaptation

FBP methods have combined geometry and appearance through shape–texture features, sample-level manifolds, landmark supervision, and visual–graph fusion (Liu et al., 2017b; Chen et al., 2023; Xiao et al., 2021; Wang et al., 2022; Peng et al., 2024; Wang et al., 2026). The multi-graph manifold method relates face samples rather than landmarks within a face (Wang et al., 2022). GPNet uses landmark prediction as supervision; DeepGeoFusion directly fuses visual and landmark-graph features for beauty regression (Peng et al., 2024; Wang et al., 2026). Thus geometry can enter as an input, a training signal, or a feature fused with appearance. Our geometry features instead condition a correction to the image embedding before ordinal prompt scoring.

CLIP-Adapter learns bottleneck residual transformations of frozen CLIP features and blends them with the original representation (Gao et al., 2021). MATeR maps image and text embeddings to tangent spaces at their respective modality centers for training-free few-shot retrieval (Long et al., 2023). Both residual adaptation and tangent-space representations therefore precede our method. Our adapter combines an image-only term with dense-geometry-conditioned interactions, projects the correction at each image embedding, bounds its tangent displacement, and renormalizes before the fixed ordinal scorer. Its conditioning and local bound distinguish it from CLIP-Adapter's residual blending, while its per-image update and scoring task differ from MATeR's modality-centered retrieval. These are design differences; the reported FBP comparisons evaluate the complete predictor rather than isolate the geometry branch's effect.

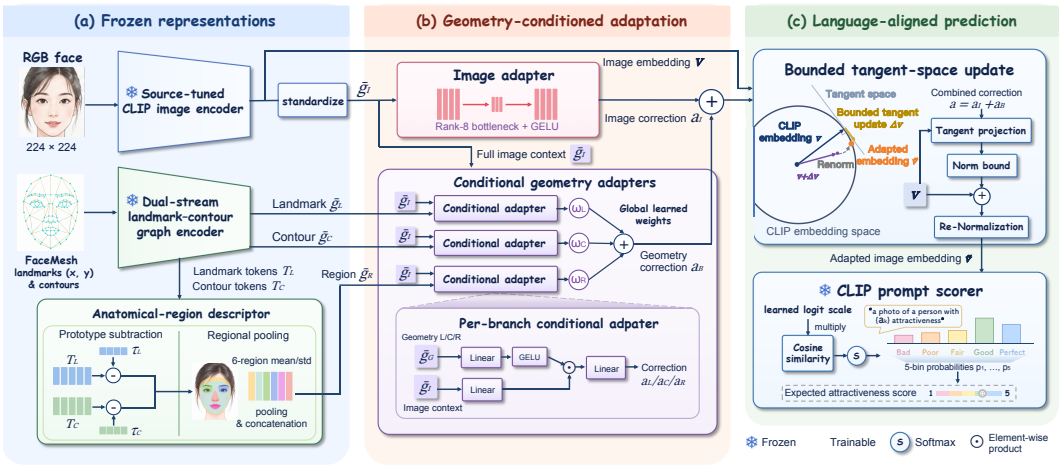

Figure 2: Final adapter stage after separate source-domain training of the appearance and geometry encoders. Equations 6–11 define the operations; the output is $\tilde{\mathbf{v}}$ only after the bounded update and normalization.

3 Methodology

3.1 Problem Formulation and Framework Overview

Let x be a face image, $L \in \mathbb{R}^{468\times 2}$ its FaceMesh coordinates, and $y \in [1, 5]$ its mean attractiveness rating. Given source data $\mathcal{D}_s$, we learn a predictor for both the source and an unseen target domain $\mathcal{D}_t$ without using target samples for training or calibration. Training has three source-domain stages, each using every image in its training split: tune the CLIP visual encoder and logit scale; train a dual-stream geometry encoder, two ordinal heads, and their probability-fusion weights from scratch; then freeze both encoders and train only the final adapter and its branch weights. The stages share the ordinal-loss components but use the stage-specific objectives in Section 3.5. Figure 2 shows the final stage. The adapter uses landmark, contour, and anatomical-region features to condition a low-rank correction of the CLIP image embedding; learned branch weights, tangent projection, norm bounding, and re-normalization constrain the update before the unchanged prompt-based scorer.

3.2 Language-Aligned Appearance Representation

Following FPEM (Li et al., 2025), we tune the CLIP ViT-B/16 visual encoder (Radford et al., 2021) and logit scale on source-training data while fixing the text encoder and five ordered prompts, then freeze the scorer before adapter training. The image embedding is

$$\mathbf{v} = \frac{f_I(x)}{\|f_I(x)\|_2} \in \mathbb{R}^{512}. \tag{1}$$

For quality terms $\mathcal{A} = \{\texttt{bad}, \texttt{poor}, \texttt{fair}, \texttt{good}, \texttt{perfect}\}$, let $\mathbf{t}_k$ denote the normalized embedding of "a photo of a person with a_k attractiveness." With the learned, subsequently frozen logit scale α, the base prediction is

$$z_k^{(0)} = \alpha \mathbf{v}^\top \mathbf{t}_k, \qquad p_k^{(0)} = \frac{\exp z_k^{(0)}}{\sum_{j=1}^{5} \exp z_j^{(0)}}, \qquad q^{(0)} = \sum_{k=1}^{5} k p_k^{(0)}. \tag{2}$$

The adapter modifies only $\mathbf{v}$; the text embeddings, logit scale, and scoring rule remain fixed.

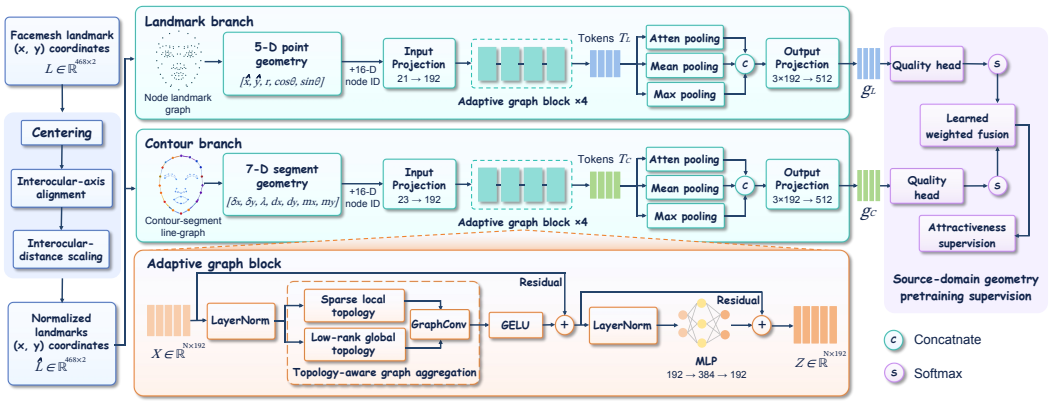

Figure 3: Dual-stream geometry encoder. Landmark and contour graphs combine sparse anatomical edges with low-rank global topology; multi-statistic pooling produces the frozen geometry representations.

3.3 Topology-Aware Geometric Representation

3.3.1 Landmark Normalization and Dual-Graph Encoding

Following LBAGCN (Zhao et al., 2024), we encode geometry with landmark and contour streams. The crop-aligned landmarks are centered, eye-aligned, and scaled by interocular distance (Kartynnik et al., 2019). The landmark graph has $N_L = 468$ point nodes; the contour graph has $N_C = 124$ segment nodes, with local neighborhoods determined by facial topology. Appendix A specifies the coordinate transform, graph edges, and node descriptors.

Each stream projects node descriptors and learned node identifiers to 192 dimensions, then applies four residual graph blocks. For $\mathbf{X} \in \mathbb{R}^{N\times 192}$, local output $\mathbf{L} \in \mathbb{R}^{N\times 192}$ aggregates each node and its fixed anatomical neighbors with learned edge weights. The global path first sends the N nodes to $R = 16$ summaries and then returns them to nodes:

$$\mathrm{GraphConv}(\mathbf{X}) = \mathbf{L} + \mathrm{sigmoid}(\gamma)\mathbf{G} + \mathbf{b}, \qquad \mathbf{G} = CS(\mathbf{X}W_{\mathrm{g}}), \quad R = 16. \tag{3}$$

Here $C \in \mathbb{R}^{N\times R}$ contains receiver weights normalized over summaries, $S \in \mathbb{R}^{R\times N}$ contains sender weights normalized over nodes, $W_{\mathrm{g}} \in \mathbb{R}^{192\times 192}$, and $\mathbf{b} \in \mathbb{R}^{192}$ is broadcast over nodes. Concatenated attention, mean, and maximum pooling produces branch features $\mathbf{g}_L, \mathbf{g}_C \in \mathbb{R}^{512}$ and 192-dimensional node tokens $\mathbf{T}_L, \mathbf{T}_C$.

3.3.2 Source Calibration and Node-Centered Geometry

Using all source-training samples $\mathcal{V}_s$, whose landmark extraction succeeded in the reported experiments, we subtract the source mean of each landmark or contour node token:

$$\boldsymbol{\tau}_{b,i} = \frac{1}{|\mathcal{V}_s|}\sum_{n\in\mathcal{V}_s} \mathbf{T}_{b,i}^{(n)}, \qquad \widetilde{\mathbf{T}}_{b,i}^{(n)} = \mathbf{T}_{b,i}^{(n)} - \boldsymbol{\tau}_{b,i}, \quad b \in \{L, C\}. \tag{4}$$

We group the centered tokens into six anatomical regions and concatenate landmark and contour means and standard deviations into $\mathbf{g}_R \in \mathbb{R}^{4608}$ (6 regions $\times 4$ statistics $\times 192$ dimensions). Appendix A gives the complete landmark indices and region-pooling details. The source-node subtraction applies only to tokens used for regional pooling; the global landmark and contour pooled branches $\mathbf{g}_L, \mathbf{g}_C$ use their original tokens. Centering removes each node's source mean, not individual identity or demographic information. For each adapter input $b \in \{I, L, C, R\}$, source-training statistics standardize and clip the feature:

$$\overline{\mathbf{g}}_b^{(n)} = \mathrm{clamp}\left(\frac{\mathbf{g}_b^{(n)} - \boldsymbol{\mu}_b}{\max(\boldsymbol{\sigma}_b, \epsilon\mathbf{1})}, -5, 5\right), \tag{5}$$

Here $\mathbf{g}_I = \mathbf{v}$, and the coordinate-wise maximum uses a positive numerical constant ϵ. We calculate each final feature's coordinate-wise mean and standard deviation from the source-training split after regional pooling, then apply this transform and the $[-5, 5]$ clipping to image, landmark, contour, and region adapter inputs. The node means and feature statistics remain fixed during source-domain checkpoint evaluation and target inference.

3.4 Asymmetric Geometry-Conditioned Tangent Adaptation

3.4.1 Low-rank image–geometry interaction

An image-only head produces

$$\mathbf{a}_I = W_{I,2}\ \mathrm{GELU}\left(W_{I,1}\overline{\mathbf{g}}_I\right). \tag{6}$$

while each geometry branch uses multiplicative image conditioning:

$$\mathbf{a}_b = W_{b,2}\left[\mathrm{GELU}(W_{b,1}\overline{\mathbf{g}}_b) \odot U_b\overline{\mathbf{g}}_I\right]. \tag{7}$$

The bias-free heads have rank eight, and learned normalized weights combine their outputs:

$$\mathbf{a} = \mathbf{a}_I + \sum_{b\in\{L,C,R\}} \omega_b \mathbf{a}_b, \qquad \omega_b = \frac{\exp(\gamma_b)}{\sum_{j\in\{L,C,R\}}\exp(\gamma_j)}. \tag{8}$$

Each $\mathbf{a}_b$ depends on its sample's geometry and image feature, whereas the learned ω_b are global parameters fixed after training; geometry does not produce an independent final score.

3.4.2 Tangent-space projection and bounded update

We remove the component of $\mathbf{a}$ parallel to $\mathbf{v}$:

$$\mathbf{u} = \mathbf{a} - \mathbf{v}(\mathbf{v}^\top\mathbf{a}) = (I - \mathbf{v}\mathbf{v}^\top)\mathbf{a}, \qquad \mathbf{v}^\top\mathbf{u} = 0. \tag{9}$$

We then bound and re-normalize the update:

$$\Delta\mathbf{v} = \mathbf{u}\min\left(1, \frac{\rho}{\max(\|\mathbf{u}\|_2, \epsilon)}\right), \qquad \widetilde{\mathbf{v}} = \frac{\mathbf{v}+\Delta\mathbf{v}}{\|\mathbf{v}+\Delta\mathbf{v}\|_2}, \qquad \rho = 0.01. \tag{10}$$

Because $\mathbf{v}^\top\Delta\mathbf{v} = 0$, the angle satisfies $\theta = \arctan\|\Delta\mathbf{v}\|_2 \le \arctan(0.01) \approx 0.573°$. This bound limits the embedding displacement; it does not preserve every text similarity or prediction ordering. The original text embeddings score $\widetilde{\mathbf{v}}$:

$$z_k = \alpha\widetilde{\mathbf{v}}^\top\mathbf{t}_k, \qquad p_k = \frac{\exp z_k}{\sum_{j=1}^5 \exp z_j}, \qquad q = \sum_{k=1}^5 k p_k. \tag{11}$$

Zero-initialized output projections make the initial model identical to the frozen source-tuned scorer. Together, source-only calibration and the bounded tangent update restrict feature-scale and alignment shifts during transfer.

3.5 Training Objectives for Ordinal Prediction and Residual Adaptation

We linearly map each dataset's ratings to $[1, 5]$ and use Gaussian-smoothed targets $\pi(y)$ over five ordinal bins. All three stages use the same four ordinal-loss components for a five-bin prediction p with expected score $q(p) = \sum_{k=1}^5 kp_k$, but the final adapter stage assigns them different weights:

$$\mathcal{L}^{\boldsymbol{\lambda}}_{\mathrm{ord}}(p, y) = \lambda_s\mathcal{L}_{\mathrm{score}}(q(p), y) + \lambda_d\mathcal{L}_{\mathrm{dist}}(p, \pi(y)) + \lambda_e\mathcal{L}_{\mathrm{cdf}}(p, \pi(y)) + \lambda_r\mathcal{L}_{\mathrm{rank}}(q(p), y). \tag{12}$$

The four terms measure score error, distribution error, CDF distance, and pairwise ranking error, respectively. The CLIP scorer and the geometry encoder use the ordinal loss to learn the source-domain appearance and geometry representations; the geometry stage supervises its landmark and contour heads as well as their fused probability. Appendix C specifies the smoothed targets, component definitions, pair-selection rule, and stage-specific weights.

For $\Delta q_i = q_i - q_i^{(0)}$ and residual target $r_i = y_i - q_i^{(0)}$, the final adapter adds residual ranking and correction penalties:

$$\mathcal{L}_{\text{res}} = 0.25\mathcal{L}_{\text{rank}}(\Delta q, r) + 0.2\left(\frac{1}{B}\sum_{i=1}^{B}\Delta q_i\right)^2 + \frac{0.02}{B}\sum_{i=1}^{B}(\Delta q_i)^2. \tag{13}$$

The adapter stage optimizes

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ord}}^{\boldsymbol{\lambda}^{\text{ada}}}(p, y) + \mathcal{L}_{\text{res}}. \tag{14}$$

The ordinal term uses adapter-specific weights given in Appendix C.

4 Experiments

4.1 Datasets and Evaluation Protocols

Table 1 gives the dataset sizes, rating scales, and train/test proportions used by FPEM (Li et al., 2025). LiveBeauty contains live-streaming Asian faces, MEBeauty ethnically diverse in-the-wild faces, and SCUT-FBP5500 Asian and Caucasian frontal faces (Lebedeva et al., 2022; Liang et al., 2018). MEBeauty labels are mapped by $y_{[1,5]} = 1 + 4(y_{[1,10]} - 1)/9$. For cross-dataset evaluation, the complete pipeline is trained and selected using LiveBeauty source data and applied unchanged to the evaluated MEBeauty and SCUT-FBP5500 images; target images and statistics do not enter training or selection. At every training stage, checkpoint selection uses the corresponding source-domain test partition and jointly considers SRCC, PLCC, and KRCC. For within-dataset results, that same test partition is also used for the reported metrics, so those results are conditional on test-partition checkpoint selection. The FPEM rows in our result tables are published values, while our source-tuned CLIP base and full-model rows come from our evaluations. We report SRCC, PLCC, and KRCC, where higher is better.

Table 1: Dataset sizes, original rating scales, reported annotator totals, and within-dataset train/test proportions. Annotator totals do not imply that each face received that many ratings.

Dataset	Images	Scale	Annotators	Train/test
LiveBeauty	10,000	$[1, 5]$	20	90/10
MEBeauty	2,550	$[1, 10]$	300	80/20
SCUT-FBP5500	5,500	$[1, 5]$	60	60/40

4.2 Preprocessing and Implementation Details

Faces are cropped by their bounding box and aspect-preserving resized to a zero-padded 224×224 canvas; MediaPipe supplies 468 FaceMesh points (Lugaresi et al., 2019; Kartynnik et al., 2019). The same crop, resize, and padding transform is applied to the landmark coordinates before the normalization in Section 3.3. Landmark extraction succeeded for every sample in the reported experiments, and all images in each source-training split were used at every stage. We sequentially train the CLIP scorer, geometry encoder, and adapter on each experiment's source-training split; Appendix B gives their full configurations and inference handling for missing landmarks. For cross-dataset evaluation, source prototypes, calibration statistics, and checkpoint selection use LiveBeauty only; within-dataset models use their respective training splits for statistics.

4.3 Within- and Cross-Dataset Results

Within-dataset performance. Table 2 compares our full model with results reported by FPEM (Li et al., 2025) under the same train/test proportions. Our full-model results are higher on all three datasets.

Cross-dataset generalization. Table 3 separates FPEM's published values from our LiveBeauty-trained CLIP base and full model. On both MEBeauty and SCUT-FBP5500, the full model exceeds the reported FPEM results on all three correlation metrics without target-domain adaptation. For

Table 2: Within-dataset comparison at the reported train/test proportions. Best and second-best results are **bold** and underlined.

Method	LiveBeauty			MEBeauty			SCUT-FBP5500		
	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑
ViT-B	0.897	0.847	0.733	0.620	0.659	0.443	0.869	0.880	0.686
ResNeXt-50	0.907	0.879	0.746	0.700	0.743	0.516	0.899	0.911	0.727
Inception-V3	0.906	0.880	0.745	0.676	0.727	0.492	0.905	0.918	0.736
AVA-MLSP	0.833	0.815	0.650	0.602	0.643	0.431	0.764	0.794	0.566
TANet	0.891	0.858	0.728	0.651	0.690	0.467	0.870	0.882	0.686
Dele-Trans	0.888	0.841	0.721	0.611	0.651	0.440	0.870	0.860	0.682
EAT	0.893	0.867	0.728	0.640	0.697	0.462	0.853	0.872	0.665
ComboNet	0.902	0.871	0.741	0.651	0.692	0.472	0.897	0.907	0.725
2D-FAP	0.896	0.854	0.733	0.690	0.719	0.506	0.903	0.915	0.734
REX-INCEP	0.908	0.884	0.750	0.702	0.739	0.514	0.907	0.917	0.739
CNN-ER	0.913	0.888	0.757	0.706	0.748	0.519	0.910	0.922	0.745
MEBeauty	0.826	0.796	0.643	0.649	0.677	0.467	0.799	0.802	0.599
FPEM	0.924	0.896	0.773	0.783	0.795	0.593	0.930	0.932	0.774
Ours	**0.932**	**0.920**	**0.783**	**0.790**	**0.806**	**0.599**	**0.937**	**0.941**	**0.788**
Improvement (%)	+0.87%	+2.68%	+1.29%	+0.89%	+1.38%	+1.01%	+0.75%	+0.97%	+1.81%

Table 3: Cross-dataset generalization from LiveBeauty. Best and second-best results are **bold** and underlined, including ties.

Method	MEBeauty			SCUT-FBP5500		
	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑
ViT-B	0.573	0.583	0.402	0.553	0.548	0.394
ResNeXt-50	0.535	0.546	0.372	0.478	0.467	0.335
Inception-V3	0.575	0.581	0.404	0.484	0.480	0.339
AVA-MLSP	0.587	0.597	0.414	0.400	0.412	0.275
TANet	0.551	0.529	0.384	0.401	0.411	0.277
Dele-Trans	0.576	0.572	0.402	0.480	0.476	0.335
EAT	0.603	0.624	0.427	0.484	0.466	0.338
ComboNet	0.540	0.548	0.375	0.384	0.375	0.264
2D-FAP	0.556	0.489	0.390	0.444	0.445	0.306
REX-INCEP	0.588	0.592	0.413	0.473	0.471	0.332
CNN-ER	0.589	0.596	0.414	0.502	0.499	0.352
MEBeauty	0.520	0.516	0.361	0.366	0.339	0.251
FPEM	0.630	0.629	0.447	0.599	0.577	0.434
CLIP base (ours)	0.628	0.626	0.447	0.724	0.726	0.535
Ours	**0.673**	**0.665**	**0.484**	**0.740**	**0.767**	**0.547**
Improvement (%)	+6.83%	+5.72%	+8.28%	+23.54%	+32.93%	+26.04%

LiveBeauty-to-SCUT-FBP5500 transfer, its relative SRCC, PLCC, and KRCC gains over FPEM are 23.5%, 32.9%, and 26.0%, respectively. In the same setting, PLCC rises from 0.726 for our source-tuned CLIP base to 0.767 for the full model ($+5.65\%$ relative to that base).

4.4 Ablation and Fusion Analysis

We compare single-modality models and two fusion baselines under the source-only protocol. *Probability fusion* combines five-bin distributions from the CLIP and landmark/contour/region (L/C/R) branches with learned global softmax weights. *Concat MLP* concatenates the CLIP embedding with standardized L/C/R descriptors and predicts an unconstrained residual embedding before the same text scorer. The encoders are frozen and only the corresponding fusion or adapter head is trained on LiveBeauty. The CLIP base is our source-tuned scorer before the 77.8K-parameter final adapter stage; it is not a rerun of the published FPEM checkpoint.

Table 4: Ablation and fusion results on LiveBeauty in-domain and cross-domain settings. Except for zero-shot CLIP, models train on LiveBeauty; selection uses its test set.

Method	In-domain Evaluation			Cross-dataset Evaluation					
	LiveBeauty			MEBeauty			SCUT-FBP5500		
	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑
Zero-shot CLIP	0.212	0.222	0.146	0.307	0.313	0.207	0.617	0.636	0.439
Geometry only	0.633	0.631	0.458	0.392	0.391	0.266	0.316	0.301	0.213
CLIP base	0.920	0.863	0.765	0.628	0.626	0.447	0.724	0.726	0.535
Probability fusion	0.867	0.852	0.688	0.597	0.599	0.422	0.534	0.548	0.374
Concat MLP	0.902	0.886	0.735	0.567	0.566	0.396	0.698	0.699	0.509
Ours	**0.932**	**0.920**	**0.783**	**0.673**	**0.665**	**0.484**	**0.740**	**0.767**	**0.547**

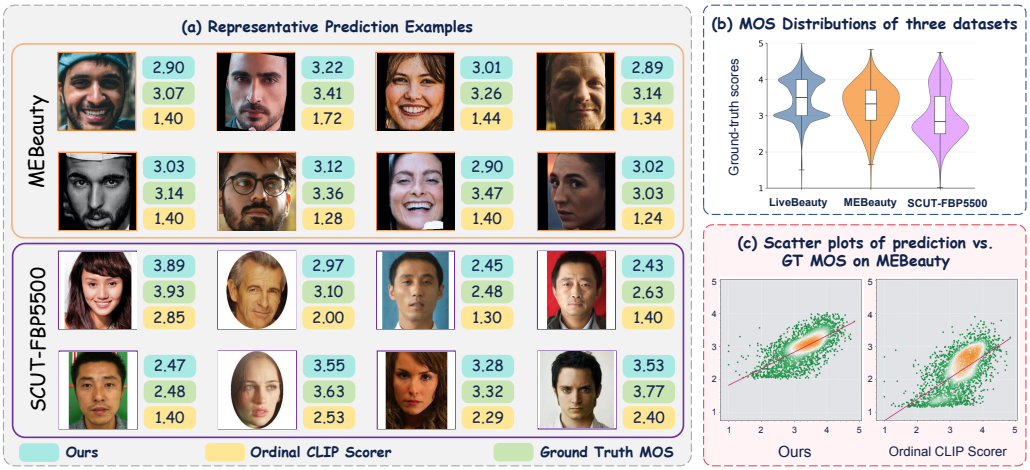

Figure 4: Qualitative results from the LiveBeauty-trained cross-dataset model. (a) Comparing illustrative samples; (b) MOS distributions of 3 datasets; (c) Predictions versus MOS on MEBeauty.

Fusion and modality ablations. Both fusion baselines trail the full model, and geometry alone trails tuned CLIP (Table 4). On MEBeauty/SCUT-FBP5500, the adapter raises PLCC from $0.626/0.726$ to $0.665/0.767$.

Remaining cross-dataset gap. Figure 4(b) shows different marginal MOS distributions after rescaling; those distributions alone do not explain correlation differences. Collection conditions, demographics, annotators, rating protocols, and label noise provide plausible sources of the remaining transfer gap (Langlois et al., 2000; Coetzee et al., 2014; Li et al., 2025; Lebedeva et al., 2022; Liang et al., 2018). The present evaluation does not measure whether the adapter reduces reliance on appearance-specific cues or establish an unavoidable performance limit.

5 Conclusion

We presented a geometry-conditioned tangent adapter that uses facial topology to refine a source-tuned CLIP scorer for cross-domain facial beauty prediction. Landmark, contour, and regional features condition a bounded update to the image embedding before the original ordinal text readout. Across three within-dataset and two cross-dataset protocols, the full model exceeds the reported FPEM results on SRCC, PLCC, and KRCC. On LiveBeauty-to-SCUT-FBP5500 transfer, PLCC reaches 0.767, compared with 0.577 for FPEM and 0.726 for our source-tuned CLIP base, without target-domain adaptation. Matched image-only adapter comparisons and tests under landmark errors and differing rating criteria would clarify geometry's contribution and robustness.

ETHICS STATEMENT

The scores modeled here summarize the judgments of particular dataset annotators; they are neither an objective measure of a person's worth nor a universal beauty standard. Landmark geometry may still reflect demographic and detector biases, and we have not established subgroup fairness. Use of identifiable face images and any displayed examples should follow the source datasets' permissions and applicable privacy requirements. The reported correlations do not justify using this system for consequential decisions about individuals without separate validation and safeguards.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT and Codex, to assist with refining aspects of the research methodology and experimental design. Additionally, these tools were used to improve the clarity and readability of the manuscript, identify potentially relevant literature, and assist in creating schematic illustrations.

We did not use generative AI to generate synthetic datasets, formulate mathematical claims or proofs, or conduct qualitative data analysis. For other tasks requiring disclosure, AI was either not used or the task was not applicable to this work.

All AI-assisted outputs were critically reviewed and verified by the authors. In particular, AI-suggested references were checked against their original sources, and the accuracy of the scientific illustrations and reported results was independently verified. The authors take full responsibility for the research contributions, experimental findings, and final content of this paper.

REFERENCES

Fares Bougourzi, Fadi Dornaika, and Abdelmalik taleb ahmed. Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowledge-Based Systems*, 242: 108246, 01 2022. doi: 10.1016/j.knosys.2022.108246.

Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. FineCLIPER: Multi-modal fine-grained CLIP for dynamic facial expression recognition with AdaptERs. In *ACM International Conference on Multimedia*, 2024. doi: 10.1145/3664647.3680827.

Huanyu Chen, Weisheng Li, Xinbo Gao, and Bin Xiao. Novel multi-feature fusion facial aesthetic analysis framework. *IEEE Transactions on Big Data*, 9(5):1302–1320, 2023. doi: 10.1109/TBDATA.2023.3255582.

Huanyu Chen, Yong Wang, Weisheng Li, and Bin Xiao. A multi-granularity facial aesthetic evaluation model based on image-text modality. *Knowledge-Based Systems*, 330:114502, 2025. doi: 10.1016/j.knosys.2025.114502.

Vinet Coetzee, Jaco M. Greeff, Ian D. Stephen, and David I. Perrett. Cross-cultural agreement in facial attractiveness preferences: The role of ethnicity and gender. *PLoS ONE*, 9(7):e99629, 2014. doi: 10.1371/journal.pone.0099629.

Yael Eisenthal, Gideon Dror, and Eytan Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2006. doi: 10.1162/089976606774841602.

Yang-Yu Fan, Shu Liu, Bo Li, Zhe Guo, Ashok Samal, Jun Wan, and Stan Z. Li. Label distribution-based facial attractiveness computation by deep residual learning. *IEEE Transactions on Multimedia*, 20(8):2196–2208, 2018. doi: 10.1109/TMM.2017.2780762.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters, 2021. URL https://arxiv.org/abs/2110.04544.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong. Predicting facial beauty without landmarks. In *Computer Vision – ECCV 2010*, volume 6316 of *Lecture Notes in Computer Science*, pp. 434–447. Springer, 2010. ISBN 978-3-642-15566-6. doi: 10.1007/978-3-642-15567-3_32.

Xing Jin, Zhihui Lai, and Zhong Jin. Learning dynamic relationships for facial expression recognition based on graph convolutional network. *IEEE Transactions on Image Processing*, 30:7143–7155, 2021. doi: 10.1109/TIP.2021.3101820.

Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2983–2991, 2015. doi: 10.1109/ICCV.2015.341.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile GPUs. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

Judith H. Langlois, Lisa Kalakanis, Adam J. Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot. Maxims or myths of beauty? a meta-analytic and theoretical review. *Psychological Bulletin*, 126(3):390–423, 2000. doi: 10.1037/0033-2909.126.3.390.

Irina Lebedeva, Yi Guo, and Fangli Ying. Mebeauty: a multi-ethnic facial beauty dataset in-the-wild. *Neural Computing and Applications*, 34:1–15, 09 2022. doi: 10.1007/s00521-021-06535-0.

Hui Li, Xiaoyu Ren, Hongjiu Yu, Ying Chen, Kai Li, L. Wang, Xiongkuo Min, Huiyu Duan, Guangtao Zhai, and Xu Liu. FPEM: Face prior enhanced facial attractiveness prediction for live videos with face retouching. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1598–1603, 2018. doi: 10.1109/ICPR.2018.8546038.

Luojun Lin, Lingyu Liang, Lianwen Jin, and Weijie Chen. Attribute-aware convolutional neural networks for facial beauty prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 847–853. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/119. URL https://doi.org/10.24963/ijcai.2019/119.

Shu Liu, Yang-Yu Fan, Zhe Guo, Ashok Samal, and Afan Ali. A landmark-based data-driven approach on 2.5d facial attractiveness computation. *Neurocomputing*, 238, 01 2017a. doi: 10.1016/j.neucom.2017.01.050.

Shu Liu, Bo Li, Yang-Yu Fan, Zhe Quo, and Ashok Samal. Facial attractiveness computation by label distribution learning with deep cnn and geometric features. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1344–1349, 2017b. doi: 10.1109/ICME.2017.8019454.

Alexander Long, Ajanthan Thalaiyasingam, and Anton van den Hengel. Modality-aware adaptation of contrastive language-image models, 2023. URL https://www.amazon.science/publications/modality-aware-adaptation-of-contrastive-language-image-models.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. URL https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

Tianhao Peng, Mu Li, Fangmei Chen, Yong Xu, and David Zhang. Geometric prior guided hybrid deep neural network for facial beauty analysis. *CAAI Transactions on Intelligence Technology*, 9 (2):467–480, 2024. doi: 10.1049/cit2.12197.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Kendra Schmid, David Marx, and Ashok Samal. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition*, 41(8):2710–2717, 2008. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2007.11.022. URL https://www.sciencedirect.com/science/article/pii/S0031320307005055.

Kunwei Wang, Xiaoyi Feng, Fadi Dornaika, Dong Huang, and Zhaoqiang Xia. A multi-graph fusion based manifold embedding for face beauty prediction. In *International Conference on Image Processing and Media Computing (ICIPMC)*, pp. 129–134, 2022. doi: 10.1109/ICIPMC55686.2022.00032.

Kunwei Wang, Yanzhi Li, Dong Huang, Junmei Feng, and Xiaoyi Feng. DeepGeoFusion: Personalized facial beauty prediction through geometric-visual fusion. *Frontiers in Computer Science*, 7, 2026. doi: 10.3389/fcomp.2025.1692523.

Qinjie Xiao, You Wu, Dinghong Wang, Yong-Liang Yang, and Xiaogang Jin. Beauty3dfacenet: Deep geometry and texture fusion for 3d facial attractiveness prediction. *Computers & Graphics*, 98:11–18, 2021. ISSN 0097-8493. doi: https://doi.org/10.1016/j.cag.2021.04.023. URL https://www.sciencedirect.com/science/article/pii/S0097849321000650.

Miao Xin, Shentong Mo, and Yuanze Lin. Eva-gcn: Head pose estimation based on graph convolutional networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1462–1471, 2021. doi: 10.1109/CVPRW53098.2021.00162.

David Zhang, Qijun Zhao, and Fangmei Chen. Quantitative analysis of human facial beauty using geometric features. *Pattern Recognition*, 44(4):940–950, 2011. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.10.013. URL https://www.sciencedirect.com/science/article/pii/S0031320310004978.

Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017. doi: 10.1109/TIP.2017.2689999.

Daqi Zhao, Jingwen Wang, Haoming Li, and Deqiang Wang. Landmark-based adaptive graph convolutional network for facial expression recognition. *IEEE Access*, 12:136088–136102, 2024. doi: 10.1109/ACCESS.2024.3463176.

Jinzhao Zhou, Xingming Zhang, Yang Liu, and Xiangyuan Lan. Facial expression recognition using spatial-temporal semantic graph network. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1961–1965, 2020. doi: 10.1109/ICIP40778.2020.9191181.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.

A Anatomical Regions and Regional Pooling

Landmark normalization and graph construction. We apply the image's bounding-box crop, aspect-preserving resize, and padding transform to its landmarks, subtract the mean of all 468 points, rotate the eye-center line to horizontal, and divide by interocular distance (Kartynnik et al., 2019). The landmark graph has undirected edges from the union of MediaPipe `FACEMESH_TESSELATION`, `FACEMESH_CONTOURS`, and `FACEMESH_IRISES`, after discarding edges with indices outside 0–467, deduplicating, and sorting. The contour graph has one node per `FACEMESH_CONTOURS` segment; two segment nodes are adjacent when their edges share an endpoint. Both local neighborhoods include the node itself. The landmark and contour streams use five- and seven-dimensional node descriptors, respectively; contour descriptors encode displacement, length, direction, and midpoint. In Equation 3, computing $S(\mathbf{X}W_{\mathrm{g}})$ before multiplication by C avoids an $N \times N$ adjacency and uses $O(NR)$ topology weights.

Anatomical regions and pooling. Table 5 lists the zero-based MediaPipe landmark indices used for the six anatomical regions. Landmark regions are disjoint; the inner-face region is the complement of the other five sets within $\{0, \ldots, 467\}$. A contour segment belongs to every region containing either of its endpoints, so contour memberships may overlap. For tokens $\mathbf{t}$ within a region, the coordinate-wise standard deviation is $\sqrt{\max(\mathbb{E}[\mathbf{t}^2] - \mathbb{E}[\mathbf{t}]^2, 0) + 10^{-6}}$.

Table 5: Zero-based MediaPipe landmark indices for the six anatomical regions. The inner-face region is the complement of the preceding five sets within $\{0, \ldots, 467\}$. A contour segment joins every region containing either endpoint.

Region	Landmark indices
Face oval	10, 21, 54, 58, 67, 93, 103, 109, 127, 132, 136, 148, 149, 150, 152, 162, 172, 176, 234, 251, 284, 288, 297, 323, 332, 338, 356, 361, 365, 377, 378, 379, 389, 397, 400, 454
Left eye and brow	249, 263, 276, 282, 283, 285, 293, 295, 296, 300, 334, 336, 362, 373, 374, 380, 381, 382, 384, 385, 386, 387, 388, 390, 398, 466
Right eye and brow	7, 33, 46, 52, 53, 55, 63, 65, 66, 70, 105, 107, 133, 144, 145, 153, 154, 155, 157, 158, 159, 160, 161, 163, 173, 246
Nose	1, 2, 4, 5, 6, 19, 45, 48, 64, 94, 97, 98, 115, 168, 195, 197, 220, 275, 278, 294, 326, 327, 344, 440
Mouth	0, 13, 14, 17, 37, 39, 40, 61, 78, 80, 81, 82, 84, 87, 88, 91, 95, 146, 178, 181, 185, 191, 267, 269, 270, 291, 308, 310, 311, 312, 314, 317, 318, 321, 324, 375, 402, 405, 409, 415
Inner face	$\{0, \ldots, 467\}$ excluding the five rows above (316 points)

B Training Configurations and Inference Handling

Table 6 gives the three source-domain training configurations. All three stages use cosine learning-rate schedules. At inference, a zero validity flag, missing landmarks, or non-finite coordinates set the geometry corrections $\mathbf{a}_L, \mathbf{a}_C, \mathbf{a}_R$ to zero; the image-only head $\mathbf{a}_I$ of the same trained adapter checkpoint still updates $\mathbf{v}$. Such samples remain in the calculation of SRCC, PLCC, KRCC, and MSE; the data loader does not filter by the validity flag. Only the zero-initialized adapter at epoch zero exactly reproduces the frozen CLIP scorer.

Table 6: Source-domain training configurations. Each stage uses the entire source-training split; its checkpoint is selected on the corresponding source-domain test partition by joint consideration of SRCC, PLCC, and KRCC. The text encoder and five ordinal prompts remain fixed throughout.

Stage	Trainable components	Epochs / batch	Optimizer and learning rate	Objective
CLIP scorer	Visual encoder and logit scale	10 / 16	PreciseAdjustAdamW, 5×10^{-6}	$\mathcal{L}_{\mathrm{CLIP}}$
Geometry encoder	Dual-stream encoder, two five-bin heads, global probability-fusion weights	70 / 16	AdamW; encoder 10^{-4}, heads 3×10^{-4}, fusion weights 10^{-3}	$\mathcal{L}_{\mathrm{geo}}$
Final adapter	Image-only and geometry-conditioned heads, adapter branch weights; both encoders frozen	12 / 16	AdamW, 10^{-4}	$\mathcal{L}_{\mathrm{total}}$

C Ordinal Loss and Stage-Specific Objectives

We form the five-bin ordinal target with a fixed Gaussian kernel over $\mathcal{K} = \{1, 2, 3, 4, 5\}$:

$$\pi_k(y) = \frac{\exp\left[-(k-y)^2/(2\sigma^2)\right]}{\sum_{j\in\mathcal{K}} \exp\left[-(j-y)^2/(2\sigma^2)\right]}, \qquad k \in \mathcal{K}, \quad \sigma = 0.55. \tag{15}$$

For batch size B, $\mathcal{L}_{\mathrm{score}}$ is the batch mean of Smooth-L1 with transition $\beta = 1$, and $\mathcal{L}_{\mathrm{dist}} = -B^{-1} \sum_{i,k} \pi_k(y_i) \log p_{ik}$. The CDF distance averages over all $B \times 5$ entries, including the final zero-difference bin:

$$\mathcal{L}_{\mathrm{cdf}}(p, \pi) = \frac{1}{5B} \sum_{i=1}^{B} \sum_{k=1}^{5} \left| \sum_{j=1}^{k} (p_{ij} - \pi_j(y_i)) \right|. \tag{16}$$

For scores s and targets r, let $\mathcal{P}(r) = \{(i, j) : i < j,\ |r_i - r_j| > 0.05\}$. The ranking term averages $\mathrm{softplus}[-\mathrm{sign}(r_i - r_j)(s_i - s_j)]$ over $|\mathcal{P}(r)|$ qualified pairs and returns zero when that set is empty.

The CLIP scorer optimizes $\mathcal{L}_{\mathrm{CLIP}} = \mathcal{L}_{\mathrm{ord}}^{\boldsymbol{\lambda}^{\mathrm{pre}}}(p^{(0)}, y)$ with $\boldsymbol{\lambda}^{\mathrm{pre}} = (1, 0.25, 0.15, 0.5)$. The geometry encoder is trained from scratch with landmark and contour five-bin heads p_L, p_C. Learned global positive weights w_L, w_C form the fused probability $p_F = w_L p_L + w_C p_C$; each head and the fusion output is supervised through its probability distribution and expected score:

$$\mathcal{L}_{\mathrm{geo}} = \frac{\mathcal{L}_{\mathrm{ord}}^{\boldsymbol{\lambda}^{\mathrm{pre}}}(p_F, y) + 0.2\mathcal{L}_{\mathrm{ord}}^{\boldsymbol{\lambda}^{\mathrm{pre}}}(p_L, y) + 0.2\mathcal{L}_{\mathrm{ord}}^{\boldsymbol{\lambda}^{\mathrm{pre}}}(p_C, y)}{1.4}. \tag{17}$$

Fusion supervision acts on the mixed probability and its expected score, rather than on an intermediate fusion embedding. The final adapter uses $\boldsymbol{\lambda}^{\mathrm{ada}} = (1, 0.05, 0.025, 0.1)$ for its ordinal term. The absolute-score and residual ranking terms use the same qualified-pair rule but different scores and targets. The preservation and counterfactual coefficients are zero in the reported training runs.