SHAPE-BAYES: BAYESIAN INFERENCE OF STRUCTURED SHAPES UNDER VISUAL AMBIGUITY

Anonymous authors
Paper under double-blind review

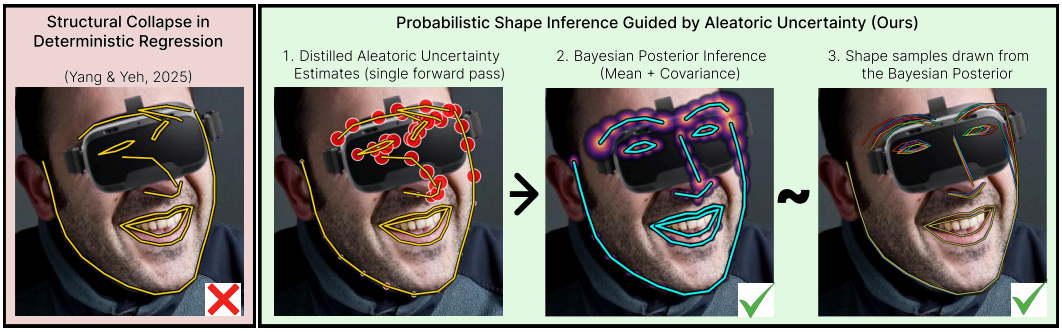

Figure 1: **Robust Bayesian shape inference under severe occlusion.** (**Left**) Deterministic regression exhibits structural collapse when forced to predict point estimates for obscured regions. (**Right**) **SHAPE-BAYES** probabilistically reconciles noisy visual evidence and a latent shape prior via aleatoric uncertainty, and infers a well-calibrated posterior. This not only guarantees structural integrity but enables the sampling of multiple plausible shape hypotheses under severe ambiguity.

ABSTRACT

Perceiving structured shapes, such as human faces, from pixels is an inherently ambiguous task in real-world conditions. Yet, shape inference is largely posed as a deterministic regression task predicting fixed spatial coordinates. We find that deterministic regression is brittle when visual evidence is ambiguous or incomplete; under severe occlusions deterministic models exhibit structural collapse, predicting incoherent shapes or reverting to generic averages. To address this, we introduce **SHAPE-BAYES**, a probabilistic framework that couples uncertainty-aware visual perception with Bayesian shape reasoning. Rather than forcing point estimates, SHAPE-BAYES dynamically weights visual evidence against geometric priors to infer a structurally valid shape posterior. Demonstrated on human face shape regression, a rigorous testbed featuring complex non-rigid deformations and strict anatomical constraints, SHAPE-BAYES comprises: (1) a base model predicting noisy landmarks alongside distilled aleatoric uncertainties; (2) a lightweight Transformer encoding these observations into an adaptive prior over a PCA shape manifold; and (3) a differentiable Bayesian solver computing closed-form posteriors by balancing the noisy predictions against this prior. By guaranteeing complete structural integrity (100% In-Distribution Rate), SHAPE-BAYES achieves an absolute improvement of up to $\sim$34% over state-of-the-art deterministic models. Simultaneously, it yields highly calibrated uncertainty bounds and reduces relative error by up to 12.5%, establishing a new state-of-the-art for robust 2D face shape regression under severe occlusion.

1 INTRODUCTION

Perceiving complex, articulated shapes is an inherently ambiguous task, particularly when visual features are occluded. Cognitive science has long posited that human vision resolves visual ambiguity through Bayesian inference (Kersten et al., 2004; Mamassian, 2006): the visual system continuously reconciles incomplete, noisy bottom-up observations with robust top-down structural priors. Classic computer vision formalized these probabilistic foundations for shape inference (Mio

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.

et al., 2004; Feldman & Singh, 2006), principles that have recently seen a resurgence in 3D generative modeling (Xu et al., 2024). In contrast, modern shape regression has largely shifted away from these Bayesian origins in favor of deterministic point-estimate paradigms. To investigate the key vulnerabilities of deterministic regression, the human face serves as a rigorous and challenging testbed. As a ubiquitous geometric structure capable of extreme rigid and non-rigid deformations, the face provides an ideal environment to study structured shape inference under severe ambiguity.

Deep deterministic regression models have driven immense progress in structural alignment tasks, achieving remarkable accuracy on clean, unoccluded targets (Jin et al., 2021; Zhou et al., 2023; Yang & Yeh, 2025). However, we find that deterministic models exhibit structural collapse under severe occlusion; when forced to predict point estimates for a hidden region lacking meaningful visual evidence, they predict structurally invalid shapes. As Figure 1 shows, they output either topologically invalid geometries or regress to a generic average shape. While recent specialized architectures attempt to address this by explicitly modeling occlusion probabilities or structural relations (Zhu et al., 2022b;a; Li et al., 2024; Chiang et al., 2025; Xiang et al., 2026), they predominantly remain deterministic, predicting a single mean shape and entirely ignoring the true distribution of probable shapes. Although some works have begun exploring uncertainty in facial alignment (Kumar et al., 2020; Xia et al., 2023), a principled framework to infer a structurally valid shape distribution under severe occlusion remains a critical missing piece.

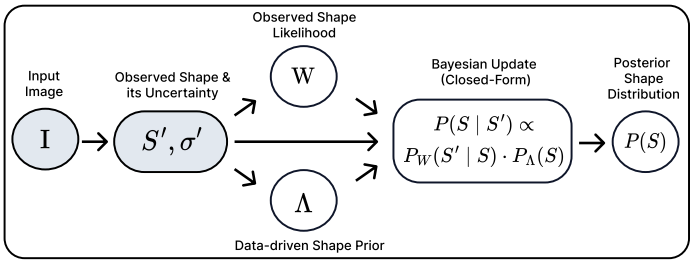

Figure 2: **Schematic overview of SHAPE-BAYES.** Given an initial noisy shape estimate, Shape-Bayes dynamically balances point-wise observation likelihoods (W) against a data-driven structural prior (Λ) via a closed-form Bayesian update to recover a calibrated posterior distribution $P(S)$. Although demonstrated only on face shape regression here, SHAPE-BAYES is topology-agnostic and applicable to any other structured shape inference.

We introduce **SHAPE-BAYES**, a Bayesian shape inference framework guided by aleatoric (data) uncertainty, as illustrated in Figure 2. The core intuition is simple: trust visual evidence where high-confidence observations exist, and rely on anatomical priors where visual evidence is ambiguous. To arbitrate between these two, we estimate distilled aleatoric uncertainty from a teacher model. This localized uncertainty acts as a spatial gate, signaling to the Bayesian update exactly which predicted coordinates are reliable. By fusing this uncertain observation with a data-driven geometric prior, SHAPE-BAYES infers a posterior distribution of coherent shapes rather than a brittle point estimates.

We evaluate SHAPE-BAYES on clean and severely occluded versions of 300W (Sagonas et al., 2013), COFW (Burgos-Artizzu et al., 2013), and WFLW (Wu et al., 2018). Across six diverse SOTA base models, it systematically reduces occluded landmark error (decreasing NME_{occ} by for HR-noSA (Yang & Yeh, 2025)) and absolute failure rates (reducing FR_{10} by 13.7% for PiPNet (Jin et al., 2021)). Crucially, SHAPE-BAYES prevents structural collapse. While deterministic models degrade under severe occlusion (e.g., In-Distribution Rate (IDR) falling to 66.6% for HR-noSA), our prior-constrained Bayesian update guarantees 100% IDR. Furthermore, it yields well-calibrated shape posteriors, lowering Expected Calibration Error (e.g. ECE from $7.56 \rightarrow 1.97$ for HR-noSA).

Our key contributions are:

- **SHAPE-BAYES Framework:** A principled Bayesian shape inference framework that leverages aleatoric uncertainty to infer full shape posteriors under visual ambiguity.
- **Aleatoric Uncertainty Distillation:** An architecture-agnostic distillation scheme that extracts predictive variance via spatial and photometric augmentations to localize occlusion.
- **Differentiable Bayesian Shape Solver:** A closed-form, differentiable solver that dynamically balances localized observation noise against a Transformer-encoded shape prior.
- **SOTA Robustness and Accuracy in Face Shape Regression:** Guarantees absolute structural integrity, improves shape accuracy across existing SOTA baselines and three challenging in-the-wild benchmarks, and yields well-calibrated posterior shape distributions.

2 SHAPE-BAYES

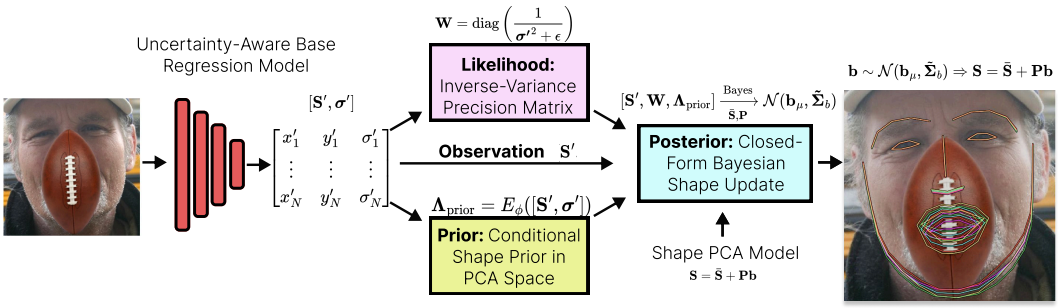

Figure 3: **SHAPE-BAYES inference flow.** (1) A base regression model predicts landmarks $\mathbf{S}'$ and uncertainties $\boldsymbol{\sigma}'$, defining the likelihood precision $\mathbf{W}$. (2) A Transformer maps these observations to an adaptive prior precision $\boldsymbol{\Lambda}_{\text{prior}}$ in PCA space. (3) A closed-form Bayesian update computes the shape posterior $\mathcal{N}(\mathbf{b}_\mu, \tilde{\boldsymbol{\Sigma}}_b)$, yielding probable shapes via sampling and projection ($\mathbf{S} = \bar{\mathbf{S}} + \mathbf{Pb}$).

Deterministic shape regressors (e.g., Yang & Yeh (2025)) exhibit structural collapse under severe occlusion (see Figure 1), predicting structurally invalid geometries when visual evidence is obscured. To move beyond these brittle point estimates, we introduce SHAPE-BAYES, a probabilistic framework that infers a structurally valid posterior shape distribution by explicitly balancing visual observations against structural priors (Figure 3). The framework operates in three distinct steps: (1) uncertainty-aware base regression, (2) an observation likelihood paired with a Transformer-encoded adaptive shape prior, and (3) exact, fully differentiable closed-form Bayesian inference.

2.1 UNCERTAINTY-AWARE SHAPE REGRESSION USING A BASE MODEL

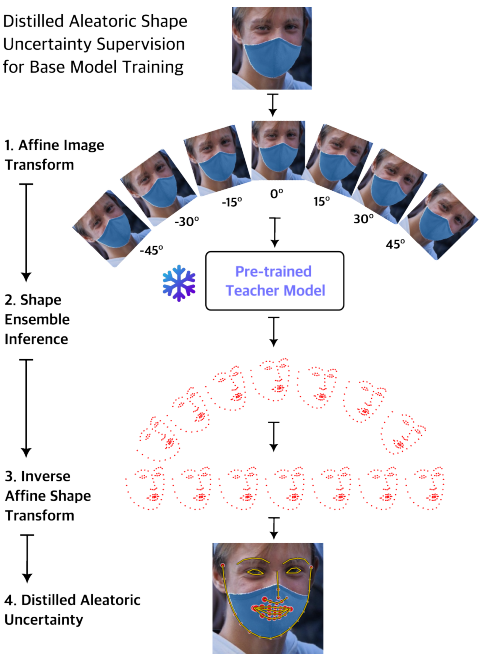

Figure 4: **Distilled Uncertainty Supervision.** Teacher predictions across affine-augmented views of an input are inverse-transformed; their empirical variance provides aleatoric targets for the base model.

To effectively arbitrate between visual evidence and prior geometric knowledge, we first estimate the landmark-wise uncertainty of the observed shape. We introduce an uncertainty-aware base regression model, Φ_θ, which estimates both the initial landmark coordinates $\mathbf{S}' \in \mathbb{R}^{N\times 2}$ and their corresponding aleatoric (data) uncertainties $\boldsymbol{\sigma}' \in \mathbb{R}^{N\times 2}$.

Supervision via Visual Feature Stability. The model Φ_θ is trained to estimate aleatoric uncertainty without manual annotations by exploiting the principle of *visual feature stability*. A visible, salient region possesses robust local features, yielding consistent coordinate predictions even if the image is subjected to spatial or photometric stress (e.g. Gaussian blur). Conversely, an occluded region lacks discriminative evidence; a deterministic model is forced to extrapolate based on weak global context. When the image is affine transformed (e.g. rotatedd) , such extrapolation may result in scattered predictions.

As shown in Fig. 4, given a training image $\mathbf{I}$, we construct M affine-augmented views $\{\mathcal{T}_i(\mathbf{I})\}_{i=1}^M$ (Appendix A). A frozen deterministic teacher Φ_T predicts shapes for each view. Applying the inverse transformation $\mathcal{T}_i^{-1}$ projects these coordinates back to the original image space, yielding hypotheses $\{\mathbf{S}_i\}_{i=1}^M \in \mathbb{R}^{M\times N\times 2}$. Taking the unaugmented prediction as the pseudo-ground truth target $\mathbf{S}'_{\text{target}}$, we compute the measurement uncertainty target as the ensemble's root mean square deviation: $\boldsymbol{\sigma}'_{\text{target}} =$

$(1/M \sum_{i=1}^{M} (\mathbf{S}_i - \mathbf{S}'_{\text{target}})^2)^{1/2}$. This establishes a heteroscedastic noise model where high local variance acts as a spatial gate, explicitly signaling regional occlusion to the Bayesian update step.

Base Model Inference in SHAPE-BAYES. The base regression model Φ_θ is trained using $\mathbf{S}'_{\text{target}}$ and $\boldsymbol{\sigma}'_{\text{target}}$ as supervision. This approach is highly flexible and architecture-agnostic, capable of leveraging any state-of-the-art architecture, from standard coordinate regressors to high-resolution heatmap regressors, provided the final layers can output the dual parameterization of coordinates and uncertainties. Importantly, during the SHAPE-BAYES inference phase, Φ_θ operates in a *single forward pass*, predicting both the initial landmarks and their corresponding aleatoric uncertainties simultaneously. This step is crucial for inferring a reliable posterior shape distribution: by outputting landmark-wise aleatoric uncertainties, the base model acts as a probabilistic gating mechanism. It explicitly dictates to the downstream Bayesian solver which spatial regions contain reliable visual evidence to anchor the shape, and which regions must be delegated to the structural shape prior.

2.2 OBSERVED SHAPE LIKELIHOOD AND CONDITIONAL SHAPE PRIOR ESTIMATION

While the predicted uncertainty $\boldsymbol{\sigma}'$ effectively flags unreliable visual evidence, correcting structural collapse requires grounding these predictions in a robust geometric prior. Although SHAPE-BAYES can accommodate any differentiable shape parameterization, for our facial alignment testbed, we constrain predictions to a classic linear PCA shape manifold (Cootes et al., 1995). This parameterizes any structurally valid shape via low-dimensional latent coefficients $\mathbf{b} \in \mathbb{R}^K$ as $\mathbf{S} = \bar{\mathbf{S}} + \mathbf{P}\mathbf{b}$, where $\bar{\mathbf{S}} \in \mathbb{R}^{2N}$ is the mean shape and $\mathbf{P} \in \mathbb{R}^{2N \times K}$ contains the principal components ($K \ll 2N$).

Observed Shape Likelihood. To ensure confident landmarks anchor the shape fit while discounting occluded regions, we define a Gaussian likelihood $\mathcal{N}(\mathbf{s}' \mid \bar{\mathbf{S}} + \mathbf{P}\mathbf{b}, \mathbf{W}^{-1})$ for the base model's shape prediction $\mathbf{s}' \in \mathbb{R}^{2N}$. Here, the diagonal precision matrix $\mathbf{W} = \text{diag}(1/(\boldsymbol{\sigma'}^2 + \epsilon))$ explicitly acts as a landmark-wise reliability weight, gating the influence of each spatial coordinate.

Transformer-Encoded Conditional Shape Prior. To overcome the limitations of static linear manifolds in handling large shape deformations, a lightweight Transformer E_ϕ dynamically adapts the shape prior to the current occlusion pattern (Fig. 3). By processing landmark coordinates and uncertainties as N tokens ($\mathbf{X} = [\mathbf{S}', \boldsymbol{\sigma}'] \in \mathbb{R}^{N \times 4}$), self-attention aggregates global geometric context, allowing confident visible regions (e.g., eyes) to deduce the structural placement of heavily occluded ones (e.g., a hidden jaw). The Transformer summarizes this context into a latent representation, which is linearly mapped to predict the diagonal elements of a prior precision matrix:

$$\boldsymbol{\Lambda}_{\text{prior}} = \text{diag}(\exp(E_\phi(\mathbf{X}))) \in \mathbb{R}^{K \times K} \tag{1}$$

Conceptually, $\boldsymbol{\Lambda}_{\text{prior}}$ quantifies the confidence of our structural prior *before* we reconcile it with the noisy shape predictions from the base model. The exponential activation ensures strict positivity, establishing a valid, dynamically adaptive shape prior $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_{\text{prior}}^{-1})$ over the latent shape coefficients.

2.3 DIFFERENTIABLE BAYESIAN SHAPE SOLVER

With the observed shape likelihood and the conditional prior established, we formulate shape inference as a Bayesian update. According to Bayes' theorem, the posterior distribution over the latent shape $p(\mathbf{b} \mid \mathbf{s}', \mathbf{X})$ is proportional to the likelihood multiplied by the prior. Because both distributions are Gaussian, the posterior is analytically exact and strictly Gaussian, $\mathcal{N}(\mathbf{b}_\mu, \tilde{\boldsymbol{\Sigma}}_b)$, defined by closed-form parameters:

$$\tilde{\boldsymbol{\Sigma}}_b^{-1} = \underbrace{\mathbf{P}^T\mathbf{W}\mathbf{P}}_{\text{Observation Precision}} + \underbrace{\boldsymbol{\Lambda}_{\text{prior}}}_{\text{Prior Precision}} \tag{2}$$

$$\mathbf{b}_\mu = \tilde{\boldsymbol{\Sigma}}_b \mathbf{P}^T \mathbf{W}(\mathbf{s}' - \bar{\mathbf{S}}) \tag{3}$$

These equations formalize the arbitration between visual evidence and structural geometry. Equation 2 demonstrates that the final posterior precision ($\tilde{\boldsymbol{\Sigma}}_b^{-1}$) is strictly the sum of information derived from the visual observation and the structural prior. Where a region is heavily occluded, its measurement precision $\mathbf{W}$ approaches zero. Consequently, the data term $\mathbf{P}^T\mathbf{W}\mathbf{P}$ contributes zero information for that spatial dimension, and the posterior shape relies entirely on the structural coherence enforced by $\boldsymbol{\Lambda}_{\text{prior}}$. Because the manifold dimension K is small (e.g., $K = 70$), computing the covariance matrix $\tilde{\boldsymbol{\Sigma}}_b$ via inversion for Equation 3 is computationally trivial and fully differentiable.

2.4 TRAINING AND OPTIMIZATION

With the trained base regression model Φ_θ frozen and the PCA model pre-computed, we optimize the Transformer encoder E_ϕ on occlusion-augmented shapes (see Appendix E). The training objective anchors absolute shape coordinates while simultaneously calibrating the predicted posterior distribution. Let $\mathbf{b}_{\text{gt}}$ be the ground truth PCA coefficients. The joint optimization objective is:

$$\mathcal{L} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{NLL}}\mathcal{L}_{\text{NLL}} \tag{4}$$

The first term is a standard Mean Squared Error (MSE) loss on the posterior mean, $\mathcal{L}_{\text{MSE}} = \|\mathbf{b}_\mu - \mathbf{b}_{\text{gt}}\|^2$, driving landmark localization accuracy. The second term is the Negative Log-Likelihood (NLL) of the ground truth shape under the predicted posterior:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2}(\mathbf{b}_{\text{gt}} - \bar{\mathbf{b}}_\mu)^T\tilde{\mathbf{\Sigma}}_b^{-1}(\mathbf{b}_{\text{gt}} - \bar{\mathbf{b}}_\mu) - \frac{1}{2}\log\det(\tilde{\mathbf{\Sigma}}_b^{-1}) \tag{5}$$

Crucially, applying a stop-gradient to the mean ($\bar{\mathbf{b}}_\mu = \text{sg}[\mathbf{b}_\mu]$) during NLL loss computation decouples optimization dynamics: MSE drives coordinate prediction (the *where*), while NLL strictly optimizes prior precision (the *how sure*), preventing the model from artificially lowering the loss by shifting only the mean shape instead of refining its variance.

3 EXPERIMENTAL SETUP

We evaluate SHAPE-BAYES on 2D face shape regression, a canonical geometric testbed exhibiting complex rigid and non-rigid deformations, under strict anatomical constraints. Our evaluation covers point-estimation accuracy, structural integrity, and posterior calibration under severe occlusions.

Datasets and Dynamic Occlusion. We evaluate on the test sets of 300W Sagonas et al. (2013), COFW Burgos-Artizzu et al. (2013), and WFLW Wu et al. (2018). To test out-of-distribution robustness without altering ground-truth shapes, we apply a dynamic occlusion protocol that procedurally overlays synthetic wearable masks and a library of 80+ real-world objects onto images. This yields clearly occluded landmark indices for evaluation. Full data synthesis details are in Appendix H.

Base Models and Aleatoric Uncertainty Distillation. To demonstrate its versatility, we integrate SHAPE-BAYES into six diverse state-of-the-art base models spanning heatmap, coordinate, and transformer-based paradigms: HR-noSA Yang & Yeh (2025), LUVLi Kumar et al. (2020), DSLPT Xia et al. (2023), ORFormer Chiang et al. (2025), STAR Zhou et al. (2023), and PiPNet Jin et al. (2021). These base models are fine-tuned to jointly predict coordinates $\mathbf{S}'$ and distilled aleatoric uncertainty $\boldsymbol{\sigma}'$ from their corresponding pre-trained teacher models (see Appendix F for details).

Implementation. SHAPE-BAYES comprises two key components: (1) a *Latent PCA Manifold* (a rank-K subspace capturing $> 99\%$ of structural variance), and (2) a *Prior Transformer* (a lightweight 5-layer encoder, $d_{\text{model}} = 64$) that maps noisy observations $[\mathbf{S}', \boldsymbol{\sigma}']$ to prior precision parameters $\mathbf{\Lambda}_{\text{prior}}$. To ground the prior in empirical noise, we train this Transformer using corrupted triplets $(\mathbf{S}_{\text{gt}}, \mathbf{S}', \boldsymbol{\sigma}')$ sampled from a base model (Appendix E). Crucially, SHAPE-BAYES adds negligible computational overhead ($< 0.2\%$ FLOPs relative to HR-noSA (Yang & Yeh, 2025)).

Evaluation Metrics. *(1) Localization Accuracy:* We report Normalized Mean Error (global NME_{all}, occluded NME_{occ}), minimum NME (minNME) across 100 posterior samples (Appendix J), Failure Rate (FR_{10}) and AUC_{10} (using the standard 10% NME threshold). *(2) Structural Integrity:* Following Lee et al. (2018), we quantify structural collapse via the In-Distribution Rate (IDR), the percentage of shape predictions falling within the 95% confidence hyperellipsoid of the learned PCA manifold:

$$\text{IDR} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\mathbb{I}\left((\mathbf{b}_i - \boldsymbol{\mu})^T\mathbf{\Sigma}_{\text{PCA}}^{-1}(\mathbf{b}_i - \boldsymbol{\mu}) \le \chi_K^2(0.95)\right) \tag{6}$$

where $\mathbf{b}_i$ is the projected latent vector, $\boldsymbol{\mu}, \mathbf{\Sigma}_{\text{PCA}}$ are training manifold statistics, and $\chi_K^2(0.95)$ is the chi-squared critical value. *(3) Probabilistic Calibration:* We evaluate posterior uncertainty bounds using Negative Log-Likelihood (NLL), Expected Calibration Error (ECE), and Spearman rank correlation (ρ) between predicted landmark variance and true Euclidean error.

4 RESULTS AND DISCUSSION

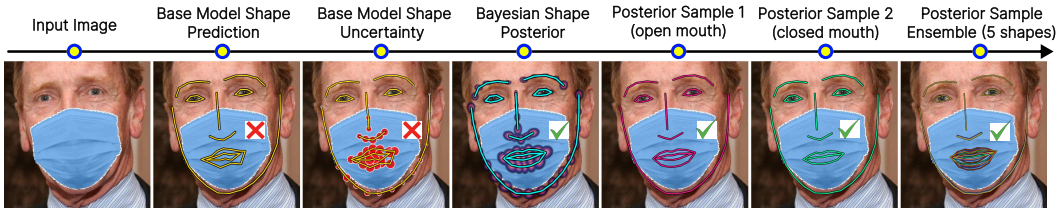

Figure 5: Qualitative illustration of the SHAPE-BAYES inference pipeline progression



Figure 6: **(Left)** Deterministic regression exhibit structural collapse. **(Right)** SHAPE-BAYES infers robust structurally valid posterior shapes.

We evaluate whether SHAPE-BAYES can recover structurally valid shapes without compromising accuracy under severe occlusions. Its core hypothesis is, conditioning of a bottom-up noisy shape prediction and its aleatoric uncertainty on a top-down shape prior yields a well-calibrated shape posteriors. We test this hypothesis along three axes: *(1) Structural integrity of the posterior mean, (2) Shape accuracy across deterministic and uncertainty-aware base models, and (3) Posterior calibration.*

Structural Integrity Evaluation. The inability of standard regression models to output structurally valid face shapes under severe occlusion is evident in their low In-Distribution Rate (IDR) scores (Tables 1 and 3). For instance, despite achieving state-of-the-art accuracy on the unoccluded test sets, HR-noSA Yang & Yeh (2025) suffers severe structural collapse under heavy dynamic occlusion, dropping to an IDR of just 66.62% on 300W. Similarly, while native uncertainty-aware baselines like LUVLi improve structural validity (82.8% on 300W), their reliance on independent, per-landmark variance without a global shape prior still allows predictions to fall outside the valid manifold. In stark contrast, SHAPE-BAYES completely resolves this vulnerability, as illustrated in Figures 5 and 6, achieving a perfect 100.00% IDR across all six evaluated base models and all three test sets. While base models suffer structural collapse on occluded 300W and WFLW, their IDR remains deceptively high on COFW due to its sparse 29-point topology. With fewer spatial constraints than dense 68- or 98-point formulations, distorted predictions on COFW rarely breach the manifold boundaries, making IDR a much stricter metric for dense topologies. Regardless of landmark density, SHAPE-BAYES guarantees a 100% IDR by strictly conditioning all predictions on the learned PCA prior.

Key Result 1: Resolution of Structural Collapse – SHAPE-BAYES guarantees structural integrity in shape predictions by inferring posterior distributions over a global shape manifold.

Shape Prediction Accuracy Evaluation. A common failure mode of PCA like prior-driven shape regularization is over-smoothing, wherein structural constraints degrade fine-grained landmark accuracy on visible regions. SHAPE-BAYES eliminates this trade-off: its well-conditioned posteriors preserve fine-grained precision on visible landmarks while consistently improving accuracy under occlusions. For HR-noSA under dynamic occlusion, SHAPE-BAYES yields systematic gains across all datasets (Table 1). On 300W, the posterior mean reduces relative NME_{occ} by 12.4% and relative NME_{all} by 6.5%, extending to a 13% error reduction via 100-sample posterior draws. This boosts reliability, dropping absolute Failure Rate by 2.4%. These gains remain consistent across all facial regions (Table 2). These results show that SHAPE-BAYES successfully resolves the tension between structural integrity and

Table 1: SHAPE-BAYES (+SB) improves accuracy and ensures structural integrity (100%IDR) under dynamic occlusion (Base: HR-noSA (Yang & Yeh, 2025)).

Data	Conf.	IDR ↑	NME_{occ} ↓	NME_{all} *(M/min)* ↓	FR ↓	AUC ↑
300W	Base	66	9.03	6.31	11.8%	41.3%
	+ SB	**100**	**7.91**	**5.90 / 5.49**	**9.4%**	**44.3%**
WFLW	Base	81	12.34	9.12	25.6%	27.9%
	+ SB	**100**	**11.35**	**8.79 / 8.36**	**23.4%**	**29.8%**
COFW	Base	96	10.28	7.40	16.0%	33.3%
	+ SB	**100**	**9.16**	**7.08 / 6.41**	**13.2%**	**35.3%**

Table 3: Generalization of SHAPE-BAYES (+SB) across base models and datasets (NME_{all}: mean / $minNME_{@100}$). Base models in gray natively output uncertainty, requiring no further training.

		300W (Dynamic Occlusion)					WFLW (Dynamic Occlusion)					COFW (Dynamic Occlusion)				
Method	**Cfg.**	**IDR↑**	**NME_{occ}↓**	**NME_{all}↓**	**FR↓**	**AUC↑**	**IDR↑**	**NME_{occ}↓**	**NME_{all}↓**	**FR↓**	**AUC↑**	**IDR↑**	**NME_{occ}↓**	**NME_{all}↓**	**FR↓**	**AUC↑**
LUVLi	Base	83	9.87	7.06	11.8	36.8	90	12.27	9.08	23.3	30.9	–	–	–	–	–
	+SB	**100**	**8.86**	**6.67/6.17**	**9.9**	**39.5**	**100**	**11.33**	**8.73/7.97**	**22.0**	**32.3**	–	–	–	–	–
DSLPT	Base	–	–	–	–	–	98	12.08	8.81	25.3	28.3	–	–	–	–	–
	+SB	–	–	–	–	–	**100**	**11.46**	**8.66/8.25**	**24.6**	**29.4**	–	–	–	–	–
ORFormer	Base	78	10.36	7.07	14.4	36.5	84	10.49	8.24	21.3	33.0	100	9.56	6.61	10.8	37.1
	+SB	**100**	**9.06**	**6.75/6.33**	**12.3**	**39.6**	**100**	**10.13**	**8.07/7.57**	**20.2**	**33.8**	**100**	**8.93**	**6.48/5.97**	**9.1**	**38.0**
STAR	Base	63	11.80	8.47	26.6	27.3	85	11.38	8.41	21.4	31.5	100	10.75	8.15	21.1	27.1
	+SB	**100**	**9.70**	**7.51/6.87**	**18.6**	**34.1**	**100**	**10.69**	**8.18/7.69**	**20.0**	**32.9**	**100**	**9.53**	**7.79/7.08**	**18.3**	**29.1**
PiPNet	Base	75	14.88	10.43	41.2	18.3	92	11.07	8.40	21.4	31.3	100	10.56	8.50	19.7	28.8
	+SB	**100**	**11.72**	**8.76/8.23**	**27.5**	**25.3**	**100**	**9.90**	**8.00/7.33**	**19.3**	**33.8**	**100**	**9.51**	**7.81/7.17**	**16.0**	**31.2**

Table 4: **Results on Original Test Sets:** HR-noSA (Yang & Yeh, 2025) as the base.

Dataset	**Config.**	**IDR↑**	**NME_{all}** *(Mean / min)*↓	**FR↓**	**AUC↑**
300W	Base	99	**2.86**	**0.1**	**71.4**
	+ SB	**100**	2.81 / **2.79**	**0.1**	71.2
WFLW	Base	98	**3.97**	**2.0**	**60.8**
	+ SB	**100**	3.94 / **3.72**	2.2	60.3
COFW	Base	**100**	**4.53**	**0.4**	**54.7**
	+ SB	**100**	4.50 / **4.23**	**0.4**	54.3

Table 5: **SOTA Comparisons on Original Test Sets.** Landmark localization errors (**NME_{all}** ↓).

Method	**300W**	**WFLW**	**COFW**
AWing (Wang et al., 2019)	3.07	4.36	4.94
LUVLi (Kumar et al., 2020)	3.23	4.37	–
ViTPose (Xu et al., 2022)	3.02	4.26	4.85
ADNet (Zhu et al., 2022a)	2.93	4.14	4.68
HIH (Zhu et al., 2022b)	3.09	4.08	4.63
SLPT (Li et al., 2024)	3.17	4.14	4.79
STAR (Zhou et al., 2023)	2.90	4.03	4.62
OccFace (Xiang et al., 2026)	2.83	3.77	4.53
HR-noSA (Yang & Yeh, 2025)	2.86	3.97	4.53
+ SHAPE-BAYES (mean)	2.81	3.94	4.50
↪ **$minNME_{@100}$**	**2.79**	**3.72**	**4.23**

fine-grained accuracy through its closed-form Bayesian solver, which employs predicted point-wise uncertainty as precision weights ($W_i = \sigma_i^{-2}$). High-confidence, visible landmarks act as strong anchors that tightly constrain the latent shape pose, whereas highly uncertain, occluded points are automatically down-weighted and guided by the conditional expectations of the geometric prior.

Cross-Architecture Generalization. SHAPE-BAYES systematically improves all six base models (Table 3). On occluded 300W, PiPNet's relative NME_{occ} drops 21.2% (with a 13.7% absolute FR reduction), while STAR reduces relative NME_{occ} by 17.8% and fully restores IDR (63.0% → 100%). *Crucially, architectures with native uncertainty heads (LUVLi, DSLPT; gray rows) require no distillation. By directly ingesting LUVLi's predicted uncertainty,* SHAPE-BAYES *lowers relative NME_{occ} by 10.2% (300W) and 7.7% (WFLW), showing its efficacy as a versatile, probabilistic extension.*

Table 2: **Part-wise NME_{occ}** on 300W (Base:HR-noSA)

Region	**Base**	**+SB**
Jaw	11.2	**10.1**
Mouth	10.2	**8.9**
Left Brow	7.6	**7.0**
Right Brow	7.0	**6.4**
Nose	6.6	**5.8**
Left Eye	5.8	**5.1**
Right Eye	5.2	**4.5**

Original Test Sets and SOTA Comparison. On clean images, predicted uncertainties vanish ($\sigma_i^2 \to 0$), causing observation precision ($W_i = \sigma_i^{-2}$) to dominate the latent prior. Consequently, SHAPE-BAYES acts as an identity mapping ($\boldsymbol{\mu}_\theta \approx \hat{\mathbf{y}}$), preserving or marginally improving base accuracy (e.g., 300W NME 2.86 → 2.81 in Table 5). Furthermore, drawing 100 posterior samples yields SOTA accuracy ($minNME_{@100}$), outperforming specialized models like OccFace and ViTPose. This confirms our learned posterior space generates highly accurate geometric hypotheses.

Key Result 2: SOTA Shape Accuracy – Trusting visual evidence where confident and shape priors where uncertain, SHAPE-BAYES improves accuracy while preserving structural integrity.

Posterior Uncertainty Calibration. Well-calibrated uncertainty estimates are vital for downstream applications where detecting a prediction failure is as critical as shape accuracy. Table 6 evalu-

ates calibration metrics across base architectures on the occluded 300W benchmark. Under out-of-distribution dynamic occlusions, most base models exhibit severe overconfidence, yielding high Expected Calibration Error (ECE) scores (e.g., 7.56 for HR-noSA and 7.10 for STAR). The notable exception is LUVLi, which achieves better baseline calibration via explicit supervision of landmark covariance. SHAPE-BAYES effectively mitigates this widespread overconfidence: analytic posterior updates consistently reduce ECE across all architectures. For instance, HR-noSA's ECE drops substantially to 1.97, while LUVLi's already low ECE further decreases to 0.86. Consistent improvements in Negative Log-Likelihood (NLL) confirm that the SHAPE-BAYES posteriors provide fundamentally more reliable confidence bounds for reliable shape inference under visual ambiguity.

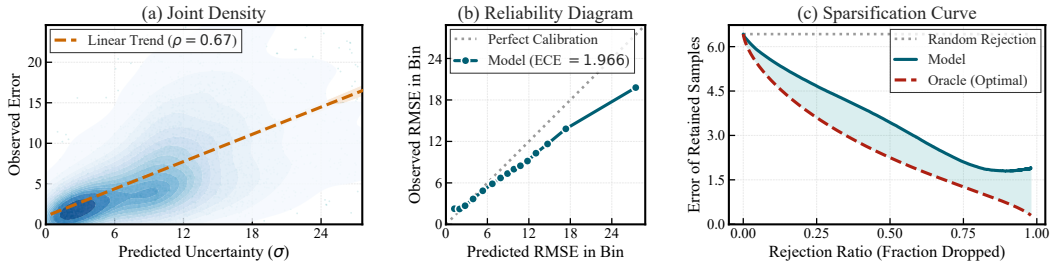

Figure 7: **Posterior Calibration.** (a) Error vs. uncertainty shows strong correlation ($\rho = 0.67$). (b) Reliability plot yields robust calibration (ECE = 1.97). (c) Sparsification curve tracks the oracle.

SHAPE-BAYES achieves this improved posterior uncertainty calibration by mapping unconstrained predictions onto a structured latent manifold, tightly aligning the predicted variance with actual empirical error. Consequently, high aleatoric uncertainty naturally propagates into a wider Gaussian posterior variance, inherently preventing overly confident mispredictions. As Figure 7 shows, the estimated aleatoric uncertainty correlates strongly with observed structural errors (Spearman $\rho = 0.67$). Moreover, the sparsification curve demonstrates that progressively filtering these uncertain predictions successfully isolates highly accurate shapes, closely tracking the oracle lower bound.

Key Result 3: Well-calibrated Shape Posterior Distributions – By propagating aleatoric uncertainty through a prior-constrained Bayesian update, SHAPE-BAYES infers well-calibrated posteriors.

Table 6: **Uncertainty Evaluation** on the occluded 300W (*denotes native uncertainty estimates from the base model.)

Method	Uncertainty	NLL↓	ECE↓
LUVLi*	Base Model*	6.08	0.90
	+SB (Posterior)	**6.07**	**0.86**
HR-noSA	Base Model	6.48	7.56
	+SB (Posterior)	**6.34**	**1.97**
ORFormer	Base Model	6.50	4.87
	+SB (Posterior)	**6.35**	**2.07**
PiPNet	Base Model	6.95	3.36
	+SB (Posterior)	**6.55**	**1.09**
STAR	Base Model	6.81	7.10
	+SB (Posterior)	**6.60**	**1.95**

Ablation Study. Table 7 details the architectural progression of SHAPE-BAYES on occluded 300W. Starting from the unconstrained *Base Deterministic Regression* baseline using HR-noSA (Occ. NME = 9.03), applying a naive *Deterministic Manifold Projection* with PCA using unweighted least-squares yields a modest error reduction (Occ. NME = 8.85). Because standard least-squares assumes isotropic, homogeneous noise ($\mathbf{\Sigma} = \sigma^2\mathbf{I}$), corrupted landmarks act as heavy outliers that distort the shape manifold and pull confident landmarks out of place.

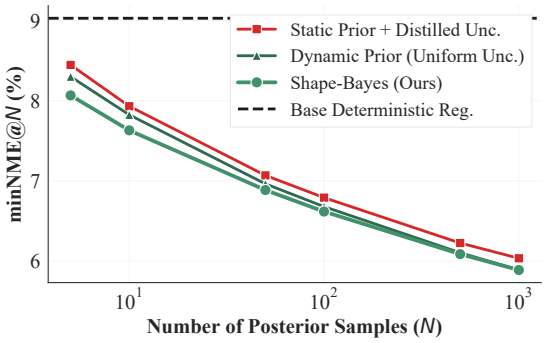

Figure 8: minNME$_{@N}$ ablation on occluded 300W

Integrating variance weights resolves this leverage issue (*Static Prior + Distilled Aleatoric Uncertainty*). Weighting landmarks by inverse variance transforms unweighted Euclidean projection into

Mahalanobis distance minimization. Severely occluded points are automatically down-weighted, allowing the structural prior to complete their positions. This variance weighting lowers Occ. NME from 8.85 to 8.08 and reduces Occ. FR from 31.57% to 25.69%. Crucially, posterior distributions enables hypothesis sampling, yielding a substantially lower Occ. minNME$_{@100}$ of 6.79 (Figure 8).

Finally, replacing the static prior with an adaptive dynamic prior yields the complete SHAPE-BAYES *(Dynamic Prior + Distilled Unc.)* framework. While a static prior can be overly rigid, the dynamic prior adaptively accommodates input-specific variations (e.g. head pose of face shapes). As Table 7 shows, this unified approach achieves optimal performance across all evaluation criteria: an Occ. NME of 7.91, Occ. minNME$_{@100}$ of 6.60, Occ. FR of 24.67%, and Occ. AUC of 31.53%.

Table 7: **Ablation of SHAPE-BAYES on occluded 300W.** Rows are ordered by architectural progression. Cell shading tracks performance from high failure rate (red) to low failure rate (blue).

Architectural Configuration	Occ. NME (↓)	Occ. minNME$_{@100}$ (↓)	Occ. FR (↓)	Occ. AUC (↑)
Base Deterministic Regression	9.03	9.03	32.15%	25.81%
Deterministic Manifold Projection	8.85	8.85	31.57%	26.97%
Static Prior + Distilled Aleatoric Uncertainty	8.08	6.79	25.69%	30.61%
Dynamic Prior (w/ Uniform Uncertainty)	8.00	6.68	25.47%	30.80%
SHAPE-BAYES (Dynamic Prior + Distilled Unc.)	**7.91**	**6.60**	**24.67%**	**31.53%**

5 RELATED WORK

Probabilistic Shape Inference. Classical computer vision relied on linear Gaussian subspaces, such as Active Shape Models (Cootes et al., 1995), to enforce geometric consistency. While modern generative models (Kingma & Welling, 2014; Xu et al., 2024) capture complex priors, they require expensive sampling and struggle to analytically ingest localized uncertainty. Concurrently, PPL (Wang et al., 2026) constructs categorical priors via discrete memory banks of prototypical poses, requiring iterative refinement. In contrast, SHAPE-BAYES couples a Transformer-encoded prior with shape uncertainty, yielding an exact, single-pass Gaussian update for structured shapes.

Uncertainty Modeling in Shape Inference. While existing shape regressors parameterize coordinate uncertainty via residual likelihoods (Li et al., 2021), Dirichlet mapping (Neven et al., 2021), adaptive heatmaps (Kumar et al., 2020), or conformal sets (Jonkers et al., 2025) for passive tasks like filtering, SHAPE-BAYES actively exploits it. Paralleling recent work on transformation-equivariant variance (Tian et al., 2025), we distill affine-augmented teacher uncertainty into base regressors. Crucially, we inject this uncertainty into a closed-form Bayesian update, elevating uncertainty from a passive metric to the core mechanism arbitrating between noisy local predictions and global priors.

Shape Inference under Occlusion. Standard shape regression exhibits structural collapse under severe occlusion. While recent methods attempt explicit occlusion modeling (Kumar et al., 2020), autoregressive joint prediction (Zheng et al., 2025), or spatial prior propagation (Zheng et al., 2026), alongside relational reasoning (Zhu et al., 2022a; Xiang et al., 2026; Chiang et al., 2025), they lack tightly coupled global shape constraints. Consequently, once local feature evidence degrades, unconstrained regression or autoregressive formulations lack a mechanism to enforce global geometric bounds, causing their predictions to drift outside the valid shape manifold. In contrast, SHAPE-BAYES performs analytical Gaussian conditioning, rigorously fusing heteroscedastic observations with a conditional shape prior in closed form to guarantee structural validity by construction.

6 CONCLUSION

We presented SHAPE-BAYES, a probabilistic framework for shape inference under visual ambiguity. By unifying distilled aleatoric uncertainty with a learned conditional prior, it computes exact closed-form posteriors. This explicitly balances noisy local evidence against global structural constraints, preventing structural collapse under severe occlusion. Benchmarked across six face alignment base architectures, SHAPE-BAYES consistently reduces shape error, sharpens uncertainty calibration, and enforces structural validity, providing a principled solution for structured shape inference.

REFERENCES

Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1513–1520, 2013.

Jui-Che Chiang, Hou-Ning Hu, Bo-Syuan Hou, Chia-Yu Tseng, Yu-Lun Liu, Min-Hung Chen, and Yen-Yu Lin. Orformer: Occlusion-robust transformer for accurate facial landmark detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 784–793. IEEE, 2025.

Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

Jacob Feldman and Manish Singh. Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, 103(47):18014–18019, 2006.

Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, 2021.

Jef Jonkers, Frank Coopman, Luc Duchateau, Glenn Van Wallendael, and Sofie Van Hoecke. Reliable uncertainty quantification for 2d/3d anatomical landmark localization using multi-output conformal prediction. *arXiv preprint arXiv:2503.14106*, 2025.

Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55(1):271–304, 2004.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8233–8243. IEEE, 2020.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Jiefeng Li, Can Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Li. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11025–11034, 2021.

Yaokun Li, Guang Tan, and Chao Gou. Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. *International Journal of Computer Vision*, 132 (4):1242–1257, 2024.

Pascal Mamassian. Bayesian inference of form and shape. *Progress in Brain Research*, 154:265–270, 2006.

Washington Mio, Anuj Srivastava, and Xiuwen Liu. Learning and bayesian shape extraction for object recognition. In *European Conference on Computer Vision*, pp. 62–73. Springer, 2004.

Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Dirichlet regressive mapping for keypoint localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12467–12476, 2021.

Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 397–403, 2013.

Lin Tian, Xiaoling Hu, and Juan Eugenio Iglesias. Test-time uncertainty estimation for medical image registration via transformation equivariance. *arXiv e-prints*, pp. arXiv–2509, 2025.

Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6970–6980. IEEE, 2019.

Ziyu Wang, Shuangpeng Han, and Mengmi Zhang. Pose prior learner: Unsupervised categorical prior learning for pose estimation. In *International Conference on Learning Representations*, volume 2026, pp. 141125–141146, 2026.

Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2129–2138, 2018.

Jiahao Xia, Min Xu, Haimin Zhang, Jianguo Zhang, Wenjian Huang, Hu Cao, and Shiping Wen. Robust face alignment via inherent relation learning and uncertainty estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10358–10375, 2023.

Xinhao Xiang, Zhengxin Li, Saurav Dhakad, Theo Bancroft, Jiawei Zhang, and Weiyang Li. Occface: Unified occlusion-aware facial landmark detection with per-point visibility. In *International Conference on Pattern Recognition*, pp. 202–217. Springer, 2026.

Haiyang Xu, Yu Lei, Zeyuan Chen, Xiang Zhang, Yue Zhao, Yilin Wang, and Zhuowen Tu. Bayesian diffusion models for 3d shape reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10628–10638. IEEE, 2024.

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

Chiao-An Yang and Raymond A Yeh. Heatmap regression without soft-argmax for facial landmark detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28729–28739. IEEE, 2025.

Hongwei Zheng, Han Li, Wenrui Dai, Ziyang Zheng, Chenglin Li, Junni Zou, and Hongkai Xiong. Hipart: Hierarchical pose autoregressive transformer for occluded 3d human pose estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16807–16817. IEEE, 2025.

Jinghong Zheng, Changlong Jiang, Yang Xiao, Jiaqi Li, Haohong Kuang, Hang Xu, Ran Wang, Zhiguo Cao, Min Du, and Joey Tianyi Zhou. Pandapose: 3d human pose lifting from a single image via propagating 2d pose prior to 3d anchor space. *Advances in Neural Information Processing Systems*, 38:90147–90169, 2026.

Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15475–15484. IEEE, 2023.

Congcong Zhu, Xiaoqiang Li, Jide Li, Songmin Dai, and Weiqin Tong. Reasoning structural relation for occlusion-robust facial landmark localization. *Pattern Recognition*, 122:108325, 2022a.

Congcong Zhu, Xintong Wan, Shaorong Xie, Xiaoqiang Li, and Yinzheng Gu. Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11102–11111. IEEE, 2022b.

APPENDIX

A QUALITATIVE RESULTS ON OCCLUDED IMAGE TEST SETS

Figures 9, 10, and 11 present qualitative comparisons between Yang & Yeh (2025), current state-of-the-art deterministic regression model and its Shape-Bayes extension on dynamically occluded test images from all three datasets. In each case, the base model exhibits characteristic structural

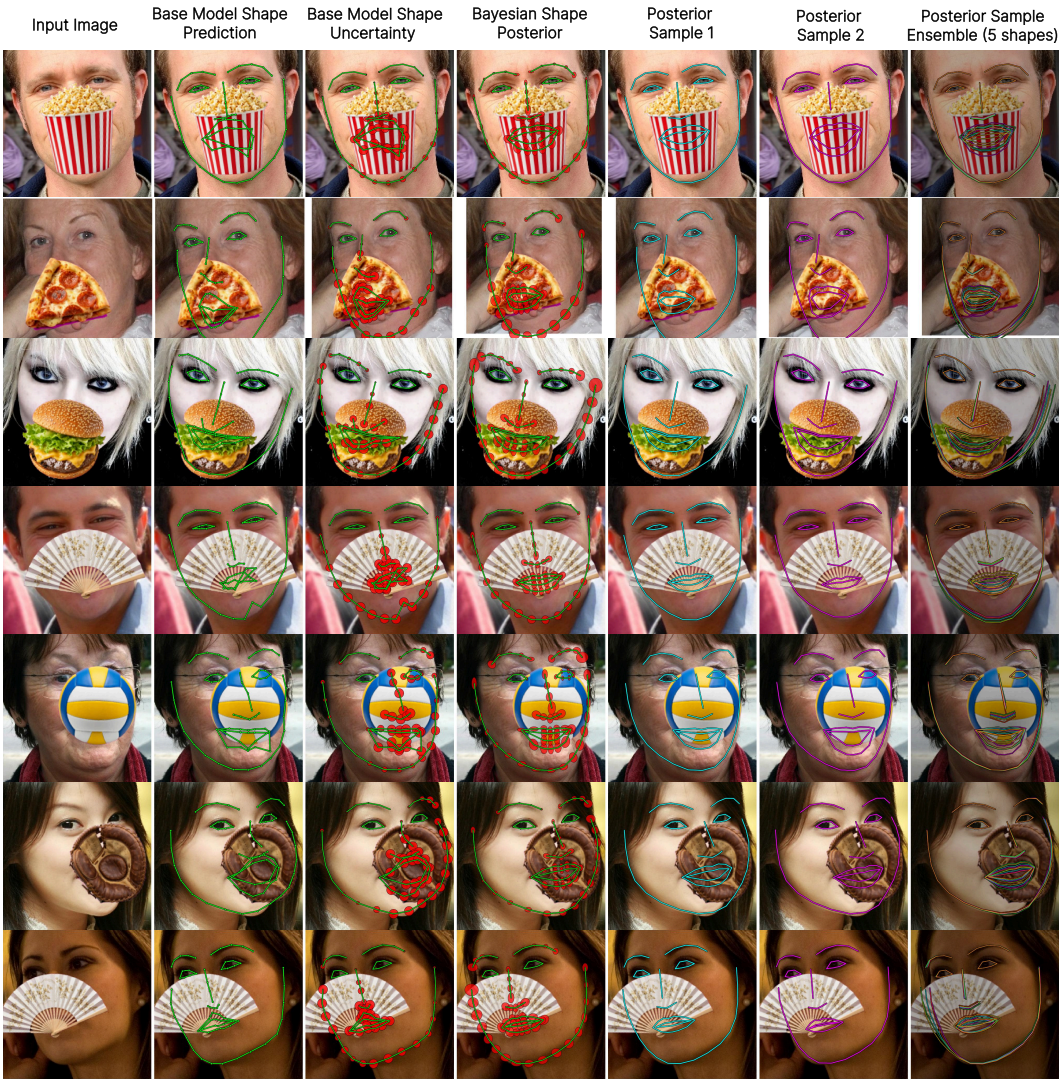

Figure 9: Qualitative results of HR-noSA Yang & Yeh (2025) as Base Model vs Shape-Bayes on **occluded 300W**.

collapse, landmarks within occluded regions drift to implausible locations, producing anatomically incoherent face shapes. Shape-Bayes corrects these failures by projecting the noisy observations onto the learned shape manifold via the closed-form Bayesian update, recovering plausible facial geometries even under severe occlusion.

B AI USAGE STATEMENT

The authors utilized Large Language Models and AI-powered assistants to support specific stages of the research and writing process. Specifically, AI coding assistants (e.g., Claude and Gemini 3 Pro) were used to aid in baseline setup and benchmarking, debugging, and visualization scripts for our experiments. Furthermore, AI-based language models were employed for proof-reading, and improving the overall clarity and readability of the manuscript. All AI-generated code and text suggestions were rigorously reviewed, edited, and validated by the authors. The authors maintain full ownership of the intellectual contributions and assume complete responsibility for the originality, accuracy, and integrity of the content, results, and software presented in this paper.

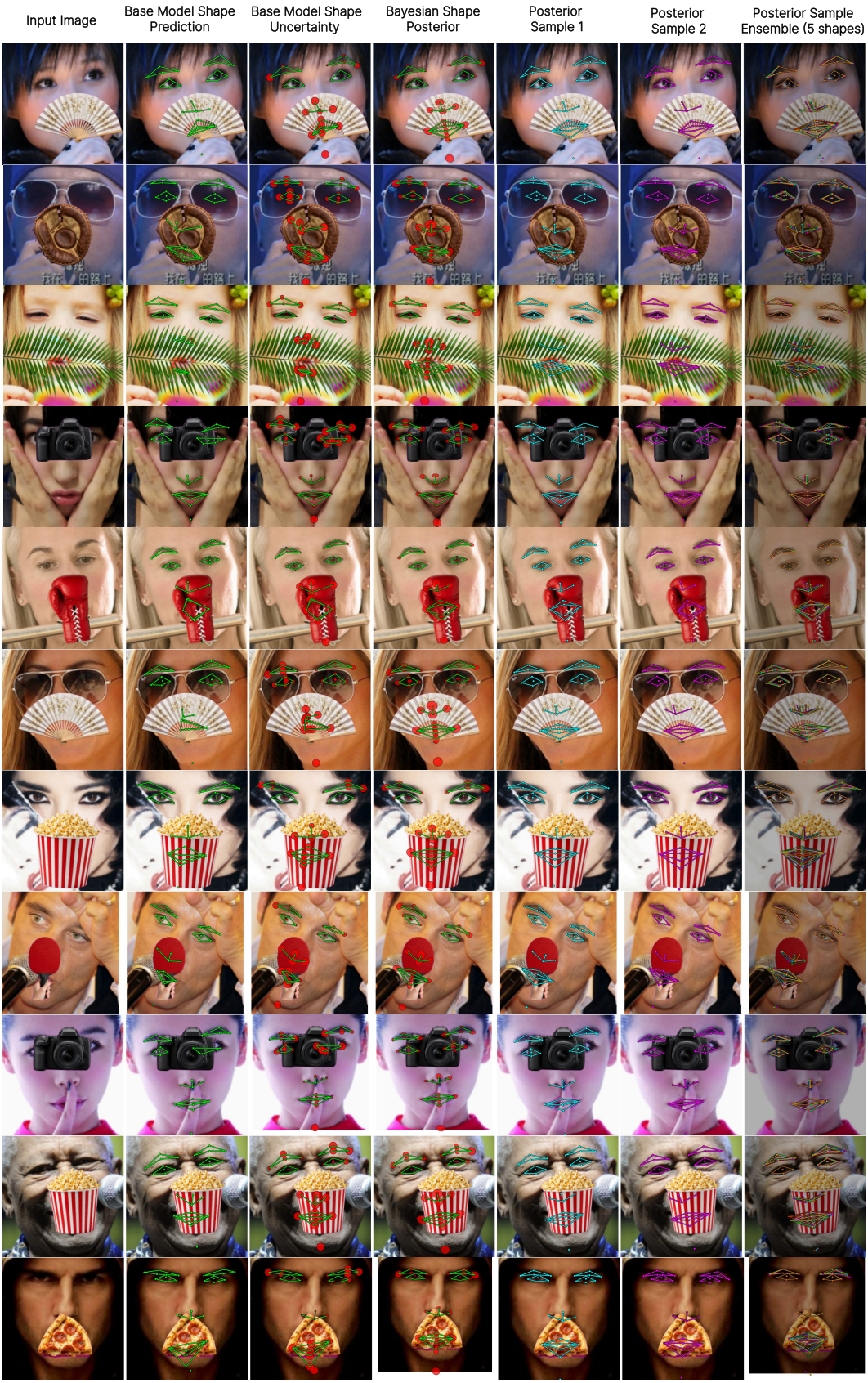

Figure 10: Qualitative results of HR-noSA Yang & Yeh (2025) as Base Model vs Shape-Bayes on **occluded COFW**

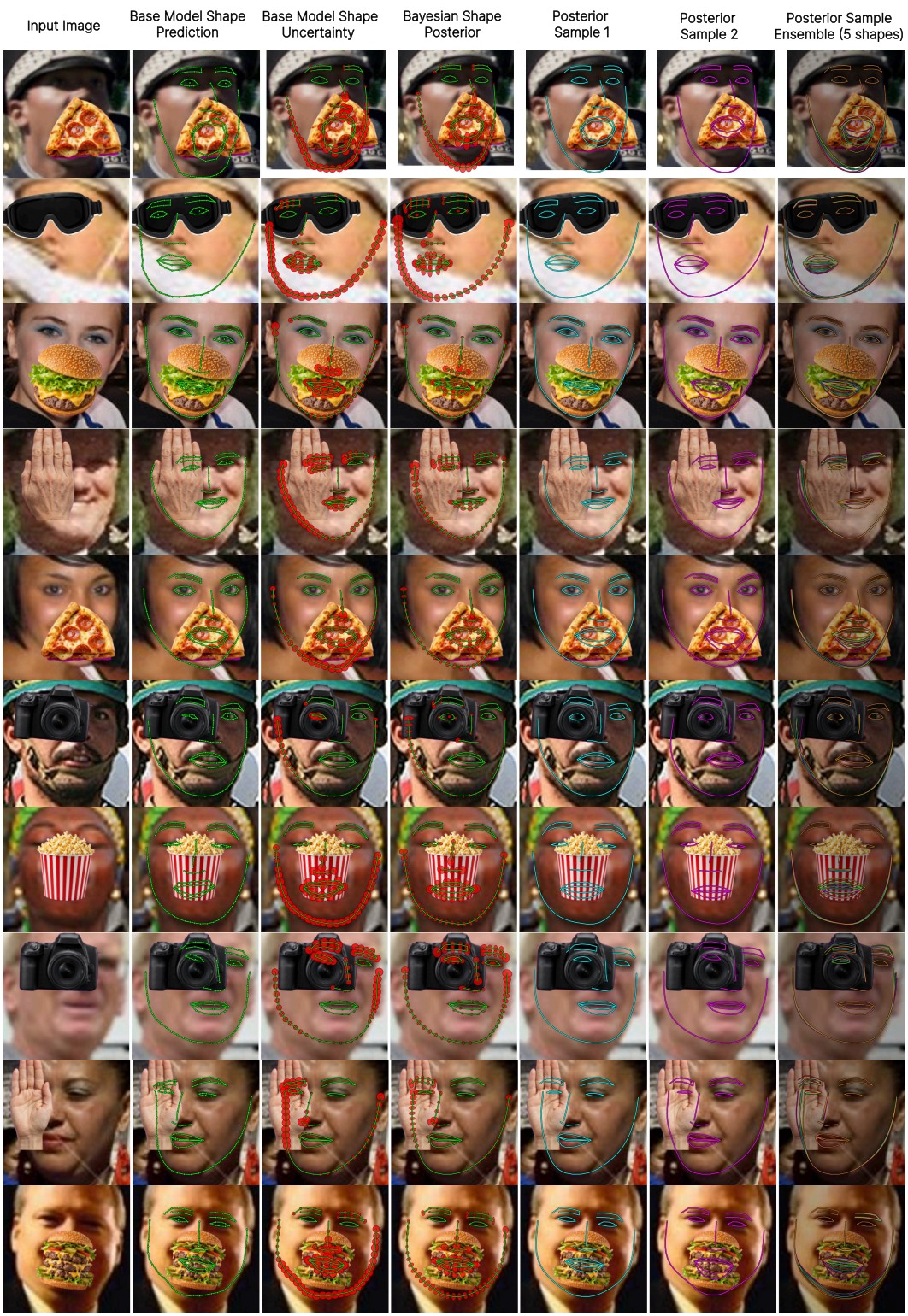

Figure 11: Qualitative results of HR-noSA Yang & Yeh (2025) as Base Model vs Shape-Bayes on **occluded WFLW**

Table 8: Region-wise shape prediction error in terms of $\text{NME}_{\text{occ}}(\downarrow)$ on **occluded 300W**.

Base Model	Config.	Jaw	R-Brow	L-Brow	Nose	L-Eye	R-Eye	Mouth
ORFormer (Chiang et al., 2025)	Base	9.0	9.5	9.8	9.5	7.1	6.6	11.1
	+SB	**8.8**	**8.8**	**9.3**	**7.8**	**6.3**	**5.7**	**9.7**
LUVLi (Kumar et al., 2020)	Base	10.7	14.1	12.9	9.0	10.2	9.3	9.3
	+SB	**9.9**	**11.3**	**11.1**	**7.4**	**8.0**	**7.7**	**8.8**
STAR (Zhou et al., 2023)	Base	13.0	10.2	12.9	10.5	9.1	7.2	12.0
	+SB	**11.5**	**9.2**	**10.6**	**7.7**	**7.3**	**6.1**	**10.2**
PIPNet (Jin et al., 2021)	Base	16.7	16.9	18.0	11.7	11.2	11.1	15.6
	+SB	**13.6**	**12.0**	**12.7**	**9.3**	**7.7**	**7.6**	**12.6**

Table 9: Region-wise shape prediction error in terms of $\text{NME}_{\text{occ}}(\downarrow)$ on **occluded WFLW**.

Base Model	Config.	Jaw	R-Brow	L-Brow	Nose	L-Eye	R-Eye	Mouth
HR-noSA Yang & Yeh (2025)	Base	11.8	11.9	11.4	11.3	9.6	10.4	12.9
	+SB	**11.5**	**11.0**	**10.9**	**9.9**	**8.9**	**9.3**	**11.8**
ORFormer (Chiang et al., 2025)	Base	11.5	11.2	11.4	9.0	9.5	8.7	10.5
	+SB	**11.0**	**10.9**	**11.2**	**8.4**	**9.1**	**8.4**	**10.2**
LUVLi (Kumar et al., 2020)	Base	11.5	12.9	13.6	11.3	11.5	11.1	12.8
	+SB	**11.1**	**12.2**	**12.7**	**10.2**	**10.5**	**10.2**	**11.8**
STAR (Zhou et al., 2023)	Base	11.4	10.4	10.2	10.6	8.9	8.6	12.6
	+SB	**11.2**	**10.0**	**9.8**	**9.6**	**8.3**	**8.2**	**11.8**
DSLPT (Xia et al., 2023)	Base	**11.3**	11.6	12.1	11.6	9.7	9.1	12.9
	+SB	11.4	**11.2**	**11.6**	**10.1**	**9.3**	**8.8**	**12.2**
PIPNet (Jin et al., 2021)	Base	12.2	14.0	13.3	9.0	10.6	11.0	11.3
	+SB	**11.2**	**11.4**	**11.3**	**8.1**	**9.3**	**9.4**	**10.3**

C Region-Wise Shape Accuracy Improvements Analysis

To verify that the Shape-Bayes manifold constraint does not introduce regional bias, for instance, improving jaw predictions at the expense of eye localization, we decompose the occluded NME into anatomically distinct facial regions. Tables 8, 9, and 10 present this breakdown for each benchmark. Across all three datasets and all base models, Shape-Bayes delivers consistent NME reductions in every facial region, confirming that the Bayesian manifold formulation regularizes coordinate estimates globally rather than trading off accuracy between components.

D Shape-Bayes Transformer: Computational Complexity Analysis

We provide a detailed breakdown of the computational cost of the Shape-Bayes Transformer encoder and the closed-form Bayesian solve, benchmarked against the StackedHGNetV1 network used in HR-noSA (Yang & Yeh, 2025). All figures are derived from the 300W configuration ($N = 68$ landmarks, $K = 60$ PCA components) and verified against PyTorch parameter counts.

D.1 Shape-Bayes Transformer Architecture

The transformer encoder maps per-landmark observation pairs $(\boldsymbol{s}_i, \boldsymbol{\sigma}_i) \in \mathbb{R}^4$ for $i = 1, \ldots, N$ to a K-dimensional log-precision vector. Its hyperparameters are: embedding dimension $d = 64$, feed-

Table 10: Region-wise shape prediction error in terms of $\text{NME}_{\text{occ}}(\downarrow)$ on **occluded COFW**.

Base Model	Config.	R-Brow	L-Brow	R-Eye	L-Eye	Nose	Mouth	Chin
HR-noSA Yang & Yeh (2025)	Base	12.0	14.4	10.0	12.1	10.7	11.1	**13.4**
	+SB	**10.6**	**11.7**	**8.5**	**8.9**	**9.1**	**9.8**	14.6
ORFormer (Chiang et al., 2025)	Base	10.1	11.6	8.0	8.8	9.3	9.9	8.7
	+SB	**9.0**	**10.4**	**7.5**	**7.9**	**8.6**	**9.5**	**8.1**
STAR (Zhou et al., 2023)	Base	11.9	14.4	9.9	12.0	10.3	10.1	13.0
	+SB	**10.7**	**12.5**	**8.6**	**9.9**	**9.0**	**9.3**	**12.8**
PIPNet (Jin et al., 2021)	Base	12.1	13.3	11.2	10.5	10.0	10.2	12.0
	+SB	**10.9**	**11.4**	**9.0**	**8.4**	**9.0**	**9.8**	**11.6**

forward dimension $d_{\text{ff}} = 128$, $L = 5$ encoder layers, and $H = 8$ attention heads (head dimension $d_h = d/H = 8$).

D.2 Parameter Count

Table 11: Shape-Bayes Transformer parameter breakdown (300W, $N = 68$, $K = 60$, $d = 64$, $d_{\text{ff}} = 128$, $L = 5$, $H = 8$). All counts verified against `torch.nn.TransformerEncoder`.

Component	Formula	Per-Unit	Subtotal
Input Projection			
$\texttt{Linear}(4 \to d)$	$4d + d$	320	320
Per Transformer Encoder Layer ($\times L = 5$)			
Multi-head self-attention			
In-projection W_{QKV}	$3d^2 + 3d$	12,480	
Out-projection W_O	$d^2 + d$	4,160	
Self-attention subtotal		16,640	
Position-wise feedforward			
$\texttt{Linear}(d \to d_{\text{ff}})$	$d \cdot d_{\text{ff}} + d_{\text{ff}}$	8,320	
$\texttt{Linear}(d_{\text{ff}} \to d)$	$d_{\text{ff}} \cdot d + d$	8,256	
Feedforward subtotal		16,576	
LayerNorm $\times 2$	$2 \times 2d$	256	
Per-layer total		**33,472**	**167,360**
Post-Encoder			
Final LayerNorm	$2d$	128	128
Precision head $\texttt{Linear}(Nd \to K)$	$N \cdot d \cdot K + K$	261,180	261,180
Total			**428,988**

Remark. The precision head dominates the parameter budget (60.9% of total), a direct consequence of flattening all N token embeddings before the final linear projection. The five transformer layers collectively contribute only 167,360 parameters (39.0%).

D.3 FLOPs Analysis

We count floating-point operations (FLOPs) using the standard convention where one multiply-accumulate equals two FLOPs, i.e., a matrix multiplication $(M, P) \times (P, Q)$ contributes $2MPQ$ FLOPs.

Summing across all stages:

where $M = 2N = 136$ is the flattened landmark dimension.

† Computing the MAP estimate $\mathbf{b}_\mu = \tilde{\mathbf{\Sigma}}_b \mathbf{P}^T \mathbf{W}(\mathbf{s}' - \bar{\mathbf{S}})$ via `torch.linalg.solve` requires only Cholesky

Table 12: Per-layer FLOPs breakdown for the Shape-Bayes Transformer ($N=68$, $d=64$, $d_{\text{ff}}=128$, $H=8$).

Operation	**Formula**	**FLOPs**
Multi-Head Self-Attention		
Q, K, V projections	$2 \cdot N \cdot d \cdot 3d$	1,671,168
Attention scores $QK^\top$	$2 \cdot N^2 \cdot d$	591,872
Attention $\times$ values	$2 \cdot N^2 \cdot d$	591,872
Output projection	$2 \cdot N \cdot d^2$	557,056
Self-attention subtotal	$2N(4d^2 + 2Nd)$	3,411,968
Position-Wise Feedforward		
$\texttt{Linear}(d \to d_{\text{ff}})$ + $\texttt{Linear}(d_{\text{ff}} \to d)$	$4 \cdot N \cdot d \cdot d_{\text{ff}}$	2,228,224
Per-layer total		**5,640,192**

Table 13: End-to-end FLOPs for the complete Shape-Bayes inference pipeline.

Stage	**Formula**	**FLOPs**
Input projection	$2 \cdot N \cdot 4 \cdot d$	34,816
Transformer encoder ($L=5$ layers)	$L \times 5.64$M	28,200,960
Precision head	$2 \cdot (Nd) \cdot K$	522,240
Neural network subtotal		28,758,016
Closed-Form Bayesian Solve (Section 2)		
$\mathbf{P}^\top \operatorname{diag}(\mathbf{w})$ scaling	KM	8,160
$(\mathbf{P}^\top \mathbf{W})\mathbf{P}$ matrix product	$2K^2M$	979,200
$(\mathbf{P}^\top \mathbf{W})(\mathbf{y} - \bar{\mathbf{s}})$	$2KM$	16,320
Cholesky factorisation + triangular solves[†]	$\frac{1}{3}K^3 + 2K^2$	79,200
Bayesian solve subtotal (MAP)		1,083,484
Full matrix inverse (for posterior sampling)[††]	$\frac{7}{3}K^3$	+432,000
Bayesian solve subtotal (MAP + sampling covariance)		1,515,484
Grand total (MAP path)		$\sim$**29.8M**
Grand total (with posterior sampling)		$\sim$**30.3M**

factorisation ($\frac{1}{3}K^3$) and two triangular solves ($2K^2$).

[††] Posterior sampling from $\mathcal{N}(\hat{\mathbf{b}}, \mathbf{\Lambda}^{-1})$ additionally requires the full matrix inverse, adding $\frac{7}{3}K^3$ FLOPs to recover $\mathbf{\Lambda}^{-1}$ via back-substitution against the identity. The dominant cost in both paths remains the $(\mathbf{P}^\top \mathbf{W})\mathbf{P}$ matrix product (979K FLOPs, $\sim$90%).

D.4 Comparison with the Base Landmark Regression Network

The base model is a StackedHGNetV1 (Yang & Yeh, 2025): a 4-stack, 4-level stacked hourglass network with CoordConv and Adaptive Attention Masks (AAM), operating on 256×256 input images. Feature maps are processed at 64×64 spatial resolution through recursive hourglass modules with bottleneck residual blocks ($C = 256$, mid-channel $= 128$).

Table 14: Computational cost comparison: Shape-Bayes module vs. StackedHGNetV1 base model.

Model	**Parameters**	**FLOPs**	**Relative Cost**
StackedHGNetV1 (base)	$\sim$13.4M	$\sim$29.1 GFLOPs	1.00$\times$
Shape-Bayes module (MAP)	429K	$\sim$29.8 MFLOPs	–
Shape-Bayes module (with sampling)	429K	$\sim$30.3 MFLOPs	–
Shape-Bayes / Base	**3.2%**	**0.10%**	$\sim$1000$\times$ cheaper

Key observations.

1. **Parameter overhead is marginal.** The Shape-Bayes Transformer adds 429K parameters (3.2% of the base network), all of which are in a separate module. The base model weights remain entirely frozen during Shape-Bayes training and inference.
2. **Compute overhead is negligible.** At ∼29.8–30.3M FLOPs (MAP to full posterior), the Shape-Bayes module constitutes approximately 0.10% of the base model's ∼29.1 GFLOPs—a ∼1000× reduction. This cost is dominated by the five-layer transformer encoder (28.2M FLOPs, 94.6%); the closed-form Bayesian solve contributes only 1.1–1.5M FLOPs (∼3.6–5.1%), of which the $\mathbf{P}^\top\mathbf{W}\mathbf{P}$ matrix product accounts for 979K FLOPs (∼90% of the solve).
3. **The bottleneck is the base model, not Shape-Bayes.** Self-attention over $N = 68$ tokens with $d = 64$ operates in the low-sequence, low-dimension regime ($N^2 d = 296{,}448$ per layer), where the quadratic attention cost is inconsequential. By contrast, the base model's hourglass modules process $64 \times 64 = 4{,}096$-token feature maps through 13 residual blocks per stack with 256-channel convolutions.
4. **Asymptotic complexity.** The transformer scales as $\mathcal{O}(L(N^2 d + Nd \cdot d_{\text{ff}}))$ in FLOPs and $\mathcal{O}(L(d^2 + d \cdot d_{\text{ff}}) + NdK)$ in parameters. For typical face alignment settings ($N \leq 98$, $K \leq 80$), both remain well below 1M FLOPs and 1M parameters, respectively, ensuring the module generalises to denser landmark configurations (e.g., WFLW with $N = 98$) without meaningful cost increases.

Wall-clock latency. On a single NVIDIA RTX 3090, the Shape-Bayes forward pass (transformer + Bayesian solve) adds $<$ 0.5 ms to the per-image inference time, which is dominated by the base model's convolutional backbone (∼12 ms). The posterior sampling step (`np.random.multivariate_normal`) is similarly negligible at < 0.1 ms per sample.

D.5 Memory Footprint

Table 15: Inference-time memory overhead of the Shape-Bayes module (FP32, batch size 1).

Component	Size	Notes
Model weights	1.64 MB	$428{,}988 \times 4$ bytes
PCA basis $\mathbf{P} \in \mathbb{R}^{M \times K}$	32.6 KB	$136 \times 60 \times 4$ bytes, stored once
Mean shape $\bar{\mathbf{s}} \in \mathbb{R}^M$	0.53 KB	136×4 bytes
Activations (peak)	∼170 KB	Dominated by $L \times N \times d$ attention maps
Precision matrix $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$	14.4 KB	Per-sample; used in `linalg.solve`
Total	$<$ **2 MB**	vs. ∼51 MB for StackedHGNetV1 weights

E Shape-Bayes Training and Optimization

The Shape-Bayes Transformer encoder is trained offline using the stochastically augmented datasets (described in Appendix I). The model processes the concatenated inputs of coordinate predictions $\mathbf{S}'$ and their associated uncertainties $\boldsymbol{\sigma}'$, yielding the diagonal prior precision matrix.

Architecture Settings. The network consists of a 5-layer Transformer encoder with an embedding dimension $d_{\text{model}} = 64$. The feed-forward network within each Transformer block has a hidden dimension of 128. Multi-head self-attention uses 8 heads, and a dropout rate of 0.1 is applied throughout to mitigate overfitting.

Optimization and Hyperparameters. The model is optimized end-to-end using the AdamW optimizer. We set the initial learning rate to 1×10^{-3} and the weight decay to 1×10^{-4}. The learning rate schedule follows a Cosine Annealing with Warm ReSTAR (Zhou et al., 2023)ts strategy, configured with $T_0 = 1$ and $T_{\text{mult}} = 2$. The model is trained for 10,000 epochs with a batch size of 64 on the 300W and COFW datasets, and for 1,000 epochs with a batch size of 32 on the WFLW dataset.

The joint loss $\mathcal{L} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{NLL}}\mathcal{L}_{\text{NLL}}$ uses $\lambda_{\text{MSE}} = 1000$ and $\lambda_{\text{NLL}} = 5 \times 10^{-4}$. The large MSE scaling compensates for gradient attenuation through the matrix inverse in the closed-form Bayesian

update, while the NLL term operates with a stop-gradient on the posterior mean to ensure decoupled optimization of shape accuracy and uncertainty quality.

Latent Manifold Dimensions. The number of principal components (K) retained for the PCA shape manifold is tailored to each dataset's native landmark topology to capture the majority of shape variance:

- **300W (68x2 shape):** $K = 60$ components.
- **WFLW (98x2 shape):** $K = 70$ components.
- **COFW (29x2 shape):** $K = 45$ components.

F Base Model Training and Implementation Details

Distilled Uncertainty Supervision Extraction for Base Model Training. For training each base network Φ_θ, we construct a Test-Time Augmentation (TTA)-based uncertainty extraction wrapper around its corresponding frozen pre-trained teacher model, that produces both the point estimate $\mathbf{S}' \in \mathbb{R}^{N\times 2}$ and a per-landmark heteroscedastic uncertainty $\boldsymbol{\sigma}' \in \mathbb{R}^{N\times 2}$. The wrapper constructs an ensemble of $M = 30$ augmented views per image: for each view $i \in \{1, \ldots, M-1\}$, stochastic spatial transformations, rotation $\theta \sim \mathcal{U}(-35°, 35°)$, isotropic scale $s \sim \mathcal{U}(1.1, 1.4)$, and translation $(t_x, t_y) \sim \mathcal{U}(-0.08, 0.08) \times (w, h)$, are applied alongside photometric perturbations (Gaussian blur with kernel $k \in \{3, 5\}$, probability 0.5) and horizontal flipping with symmetric landmark remapping (probability 0.5). The first sample ($i = 0$) is the unaugmented, deterministic prediction $\mathbf{S}' = \mathbf{S}_0$. Each transformed prediction is projected back to the original coordinate frame via the inverse affine $\mathcal{T}_i^{-1}$, and the per-landmark uncertainty is computed as:

$$\boldsymbol{\sigma}' = \sqrt{\frac{1}{M}\sum_{i=0}^{M-1}(\mathbf{S}_i - \mathbf{S}')^2} \tag{7}$$

To prevent numerical instability, $\boldsymbol{\sigma}'$ is clipped to $[10^{-6},\ 0.5 \times d_{\text{IOD}}]$, where d_{IOD} is the inter-ocular distance.

We train (a). coordinate regressors base models with heteroscedastic variance heads, and (b). heatmap regressors with anisotropic Gaussian targets. Since LUVLi (Kumar et al., 2020) and DSLPT (Xia et al., 2023) inherently output native uncertainty estimates through their probabilistic formulations, they are utilized out-of-the-box and are not fine-tuned. For the remaining four baseline models (ORFormer (Chiang et al., 2025), HR-noSA Yang & Yeh (2025), PIPNet (Jin et al., 2021), and STAR (Zhou et al., 2023)), we introduce uncertainty prediction capabilities tailored to their respective paradigms:

Heatmap Regression Base Models: For networks that predict spatial probability maps, we replace the standard isotropic Gaussian ground-truth targets with dynamic, anisotropic Gaussian heatmaps. The spatial covariance matrix of the Gaussian target for each landmark is constructed directly from the distilled empirical uncertainty $\boldsymbol{\sigma}'$. The base model is fine-tuned to minimize the divergence between its predicted heatmaps and these uncertainty-aware Gaussian targets. Consequently, the spread of the predicted heatmaps intrinsically encodes the aleatoric uncertainty, which can be extracted at inference time.

Coordinate Regression Base Models: For networks that directly regress spatial coordinates, we augment the final prediction module with a parallel heteroscedastic variance head. This auxiliary head is tasked with outputting a strictly positive variance estimate corresponding to each predicted landmark coordinate. During fine-tuning, the network minimizes a Gaussian Negative Log-Likelihood (NLL) loss, effectively forcing the variance head to match the empirical spatial covariance targets $\boldsymbol{\sigma}'$ distilled from the teacher ensemble.

G Evaluation Metric Definitions

Normalized Mean Error (NME). For a predicted shape $\hat{\mathbf{S}} \in \mathbb{R}^{N\times 2}$ and ground-truth $\mathbf{S}_{\text{gt}}$, the NME is:

$$\text{NME} = \frac{1}{|\mathcal{I}|}\sum_{n\in\mathcal{I}} \frac{\|\hat{\mathbf{s}}_n - \mathbf{s}_n^{\text{gt}}\|_2}{d_{\text{IOD}}} \times 100\% \tag{8}$$

where $\mathcal{I}$ denotes the index set over which the error is computed ($\mathcal{A}$ for all, $\mathcal{O}$ for occluded).

In-Distribution Rate (IDR). We project the predicted shape into the PCA latent space to obtain coefficients $\mathbf{b} = \mathbf{P}^T(\hat{\mathbf{s}} - \bar{\mathbf{S}})$. A shape is classified as structurally valid (*in-distribution*) if it satisfies a dual condition:

$$\text{IDR}: \quad \sum_{k=1}^{K} b_k^2 \le \chi^2_{K,0.95} \quad \text{and} \quad \|\hat{\mathbf{s}} - (\bar{\mathbf{S}} + \mathbf{P}\mathbf{b})\|_2 < 0.05 \tag{9}$$

The first condition tests whether the shape lies within the 95% confidence region of a χ^2 distribution with K degrees of freedom, and the second bounds the Mean Reconstruction Error (MRE).

Total Predictive Uncertainty. Because Shape-Bayes outputs a full posterior $\mathcal{N}(\mathbf{b}_\mu, \tilde{\mathbf{\Sigma}}_b)$, the total predictive uncertainty at each landmark combines epistemic and aleatoric components via the Law of Total Variance:

$$\boldsymbol{\sigma}^2_{\text{pred}} = \underbrace{\text{Var}_{\mathbf{b}\sim p(\mathbf{b}|\cdot)}[\hat{\mathbf{S}}(\mathbf{b})]}_{\text{Epistemic}} + \underbrace{\boldsymbol{\sigma}'^2}_{\text{Aleatoric}} \tag{10}$$

H Dynamic Occlusion: Asset Catalogue and Placement Specifications

This appendix details the two complementary occlusion synthesis pipelines used to construct the occluded test sets for evaluation (Section 3). Both pipelines operate on all three benchmarks (300W with 68-point, COFW with 29-point, and WFLW with 98-point landmark topologies), with dataset-specific anchor mappings derived from each annotation scheme.

H.1 Pipeline 1: Wearable Face Masks (MaskTheFace)

We use MaskTheFace [1] to render five categories of wearable face masks: *surgical*, *N95*, *KN95*, *cloth*, and *gas*. For each test image, one mask type is randomly sampled (uniform over the five categories when `mask_type=random`). The mask fitting procedure operates as follows:

1. **Landmark-to-Facial-Geometry Mapping.** Ground-truth landmarks are converted into a facial feature dictionary (jawline, nose bridge, eye contours, lip contours) using dataset-specific index mappings. For 300W, this uses the standard iBUG 68-point topology. For COFW (29 points), a synthetic 17-point jawline is constructed from the inter-pupillary axis, nose tip, and chin landmark by fitting an elliptical arc parameterized by the inter-ocular distance (IOD). For WFLW (98 points), native contour indices are used directly.
2. **Six-Point Affine Fit.** Six anchor points on the face (derived from the nose bridge, chin, and jawline extremes) define an affine warp that maps the canonical mask template onto the subject's facial geometry, accounting for head pose (yaw, pitch, roll).
3. **Mask Compositing.** The warped mask is alpha-composited onto the original image. A binary mask of the composited region is retained for per-landmark occlusion labelling.

H.2 Pipeline 2: Real-World Object Overlays

We curate a library of 80+ RGBA object assets spanning diverse real-world occluders. Each asset undergoes alpha-channel cleaning (binary thresholding at $\alpha = 200$, followed by 2 iterations of 3×3

[1] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication, 2020.

morphological erosion to eliminate translucent edge artefacts) and tight bounding-box cropping. The precise opaque pixel count is recorded for accurate area-based scaling.

Anatomically-Aware Placement Modes. Each asset is assigned to exactly one of eight placement modes, which determines (i) the facial anchor point for centering, (ii) the target face-area percentage range for dynamic scaling, and (iii) the rotation offset relative to the subject's head roll:

Table 16: Object overlay placement modes and representative assets.

Mode	Anchor Point	Face Area %	Representative Assets
Global Scene	Full image	100%	Chainlink fence, frost, foliage, prison bars
Helmet	Inter-ocular midpoint	110–140%	Navy/silver, white/blue helmets
Eyewear	Inter-ocular midpoint	15–35%	Goggles, VR headset, masquerade mask
Faceguard	Nose–chin midpoint	25–40%	Football helmet faceguard
Mouth	Mouth centroid	5–40%	Mug, burger, pizza, microphones, fruit
Camera	Random eye/midpoint	20–40%	DSLR, telephoto lens, TLR camera
Hand/Glove	Random facial feature	20–45%	Boxing glove, fists, open palm, mittens
Lower-Mid Blocker	Nose–mouth midpoint	15–60%	Tablet, teddy, bouquet, shipping box
Floating	Random landmark	5–50%	Football, soccer ball, balloons, palm frond

Dynamic Area Scaling. For each asset, a target face-area percentage p is uniformly sampled from the mode-specific range. The face area is estimated as an ellipse $A_{\text{face}} = \pi \cdot (w_{\text{face}}/2) \cdot (h_{\text{face}}/2)$, where w_{face} is the outer face width and $h_{\text{face}} = 1.25 \times \|\text{eye_mid} - \text{chin}\|_2$. The asset is then resized by scale factor $\sqrt{p \cdot A_{\text{face}}/A_{\text{opaque}}}$, where A_{opaque} is the pre-computed opaque pixel count, ensuring the *actual occluding area* (not bounding box) matches the target.

Occlusion Label Extraction. After compositing, the alpha channel of the overlay defines a continuous occlusion mask $\alpha(x, y) \in [0, 1]$. A landmark n is classified as occluded if $\alpha(\mathbf{s}_n^{\text{gt}}) > 0.5$, producing the per-image partition $(\mathcal{A}, \mathcal{O})$ used for decoupled evaluation.

I Data-Driven Prior: Training Augmentation Protocol

To train the Shape-Bayes Transformer encoder E_ϕ, we construct a diverse dataset of realistic $(\mathbf{S}_{\text{gt}}, \mathbf{S}', \boldsymbol{\sigma}')$ triplets by procedurally corrupting each pristine training image with 13 structured augmentation variants. All spatial parameters are anchored to the Inter-Ocular Distance for scale invariance (it is $d_{\text{IOD}} = \|\mathbf{s}_{36} - \mathbf{s}_{45}\|_2$ for 300W). The ratio of non-occluding to occluding variants (4:9) ensures the solver observes confident predictions (low $\boldsymbol{\sigma}'$) approximately 30% of the time, preventing the prior from being unconditionally strong.

I.1 Non-Occluding Photometric Variants

These variants degrade image quality without directly obscuring facial landmarks, producing training pairs where the base model remains largely confident ($\boldsymbol{\sigma}' \approx 0$):

1. **Pristine:** The unmodified training image.
2. **Brightness (Gamma Correction):** $\gamma \sim \mathcal{U}(0.55, 0.8)$. Applies a power-law LUT: $I_{\text{out}}(p) = \lfloor (I_{\text{in}}(p)/255)^\gamma \times 255 \rceil$.
3. **Contrast (Linear Adjustment):** Gain $\alpha \sim \mathcal{U}(0.5, 1.8)$, bias $\beta \sim \mathcal{U}(-30, 30)$. Applies $I_{\text{out}} = \text{clip}(\alpha \cdot I_{\text{in}} + \beta, 0, 255)$.
4. **Color Shift (HSV Perturbation):** Hue offset $\Delta h \sim \mathcal{U}(-15, 15)$ (mod 180), saturation scale $\alpha_s \sim \mathcal{U}(0.6, 1.4)$.

I.2 Occluding Structural Variants

These variants introduce spatially structured occlusions that selectively corrupt subsets of landmarks:

1. **Lower-Half Mask:** A grey rectangle $(128, 128, 128)$ with additive noise ($\mathcal{U}(0, 50)$) is placed from the nose tip (landmark 33) $\pm\mathcal{U}(-0.3, 0.3) \times d_{\text{IOD}}$ downward to the image bottom.
2. **Anatomical Sunglasses/Eyepatches:** Dark rectangles (RGB $\sim \mathcal{U}(5, 30)^3$) are fitted to bounding boxes of the eye and eyebrow landmarks (36–47) with margin $\mathcal{U}(0.2, 0.6) \times d_{\text{IOD}}$. Three asymmetric variants are sampled: both eyes ($p = 0.6$), left only ($p = 0.2$), or right only ($p = 0.2$). When both eyes are occluded, a connecting bridge strip spans the nose bridge (landmarks 27–28).
3. **Spatially Variant Blur:** A Gaussian blur with kernel size $k = \lfloor 0.8 \times d_{\text{IOD}} \rceil | 1$ is applied through a spatially varying blend mask generated from low-frequency random noise ($h/8 \times w/8$, bicubic upsampled, Gaussian smoothed, scaled $\times 1.5$ and clipped to $[0, 1]$).
4. **Extreme Random Cutout:** A rectangular patch of size $\mathcal{U}(0.5, 1.5) \times d_{\text{IOD}}$ (both axes) is placed at a random image location and filled with uniform random pixel noise $\sim \mathcal{U}(0, 255)^3$.
5. **Random Lines:** $n \sim \mathcal{U}(5, 15)$ line segments with random endpoints, random RGB color, and thickness $\max(1, \lfloor d_{\text{IOD}} \times \mathcal{U}(0.01, 0.05) \rceil)$ are drawn across the image, simulating stray hair, wires, or scratches.
6. **Vertical Half-Face Block:** A solid-color rectangle is placed from the nose bridge (landmark 30) $\pm\mathcal{U}(-0.1, 0.1) \times d_{\text{IOD}}$ to either the left or right image boundary (randomly selected with equal probability), simulating wall edges or doorframes.
7. **Additive Sensor Noise:** Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_n^2)$ with $\sigma_n \sim \mathcal{U}(10, 30)$ is added pixel-wise, simulating low-light sensor artifacts.
8. **Irregular Elliptical Blobs:** $n \sim \mathcal{U}(1, 6)$ filled ellipses with semi-axes $\mathcal{U}(0.2, 0.6) \times d_{\text{IOD}}$, random rotation, and skin-like color (RGB $\sim \mathcal{U}(180, 250)^3$) are anchored to jaw and mouth landmarks (indices 0–16 and 48–67) with positional jitter $\pm 0.5 \times d_{\text{IOD}}$, simulating hands or objects partially covering the lower face.
9. **Targeted Micro-Occlusions:** $n \sim \mathcal{U}(2, 6)$ randomly selected individual landmarks receive small circular or rectangular occlusions with radius $\mathcal{U}(0.08, 0.2) \times d_{\text{IOD}}$ and random RGB color, simulating localized debris, birthmarks, or sensor defects.

For each of the 13 variants, the frozen base model Φ_T is queried with $M = 30$ TTA passes (Section 2) to produce empirically grounded predictions $\mathbf{S}'$ and uncertainties $\boldsymbol{\sigma}'$. The per-landmark $\boldsymbol{\sigma}'$ is clipped to $[10^{-6},\ 0.5 \times d_{\text{IOD}}]$. The resulting $13 \times |\mathcal{D}_{\text{train}}|$ triplets are serialized and used for offline Shape-Bayes Transformer training.

J Diverse Posterior Sampling for minNME

The minNME metric (Section 3) requires drawing samples from the posterior $\mathbf{b} \sim \mathcal{N}(\mathbf{b}_\mu, \tilde{\boldsymbol{\Sigma}}_b)$ that systematically cover the distribution's principal modes rather than clustering near the mean. We employ a deterministic eigenvector-based sampling scheme:

1. Compute the eigendecomposition $\tilde{\boldsymbol{\Sigma}}_b = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_K)$ are eigenvalues sorted in decreasing order and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$ are the corresponding eigenvectors.
2. For each principal component $k \in \{1, \ldots, K\}$, generate two samples:
$$\mathbf{b}_{2k-1} = \mathbf{b}_\mu + 2\sqrt{\lambda_k}\,\mathbf{v}_k, \qquad \mathbf{b}_{2k} = \mathbf{b}_\mu - 2\sqrt{\lambda_k}\,\mathbf{v}_k \tag{11}$$
which places samples at $\pm 2\sigma$ along each eigenvector direction.
3. If $N_s > 2K$, additional samples are drawn via standard multivariate normal sampling.

This scheme ensures that the first $2K$ samples are located at the 95th percentile boundary along each independent axis of variation, maximizing the chance of finding a sample near the ground truth even when the posterior mean is displaced by structured error.