machine learning & knowledge extraction

MDPI

Article

Bayesian Multi-Task Facial Emotion Recognition with Reliability-Aware Uncertainty Under Controlled Facial Masking

Qiyuan Xiao and Changqin Quan *

Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan; 225x503x@stu.kobe-u.ac.jp
* Correspondence: quanchqin@gold.kobe-u.ac.jp

Abstract

Facial emotion recognition (FER) in real-world settings is limited by the semantic mismatch between discrete emotion categories and continuous Valence–Arousal–Dominance (V-A-D) dimensions and the lack of reliable uncertainty estimates under incomplete facial evidence. Existing uncertainty-aware FER studies mainly address annotation ambiguity or training-time reliability, leaving the behavior of predictive uncertainty under progressive input degradation insufficiently examined. This paper proposes BGDC (Bayesian Gaussian-mixture Distributional Consistency), a multi-task FER framework that integrates a GMM-based soft consistency module with a context-conditioned Bayesian regression head and explicitly models aleatoric and epistemic uncertainty. To evaluate predictive reliability, a controlled masking protocol is introduced to remove facial information under different spatial configurations. On FER2013-VAD, BGDC attains the highest classification accuracy of 0.6943 and the highest mean V-A-D CCC of 0.6079 among the compared configurations, and it yields a stronger epistemic uncertainty-error correspondence than MC Dropout in a single-model setting. Controlled masking further shows that the epistemic uncertainty of BGDC tracks task-relevant facial information loss rather than masking ratio alone: it rises with regression error when diagnostically important regions are removed, and it contracts when the masked region is largely task-irrelevant. Combining Bayesian uncertainty with the GMM-based distributional prior thus enables reliability-aware multi-task FER, in which controlled masking serves as a diagnostic intervention rather than as a benchmark of accuracy degradation alone.

Keywords: dimensional emotion recognition; multi-task learning; V-A-D regression; reliability-aware prediction; controlled facial masking

Check for updates

Academic Editor: Feiping Nie

Received: 18 May 2026
Revised: 11 June 2026
Accepted: 22 June 2026
Published: 25 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Affective computing aims to equip machines with the ability to perceive, represent, and respond to human emotions, and it has become a foundational component of intelligent human–computer interaction systems [1,2]. Within this broad agenda, facial expression recognition (FER) remains one of the most intensively studied visual channels because facial behavior provides rich and direct evidence of affective state. At the same time, the field is moving beyond single-task categorical recognition toward multi-task formulations that jointly model discrete emotion classes and continuous affective attributes, since these two forms of supervision describe complementary aspects of emotional behavior [2,3]. Such a formulation is attractive not only because it yields richer emotion representations but also because it enables the sharing of cross-task structure that can improve both recognition and regression performance [3–5].

A first fundamental challenge in this setting is the semantic coupling between categorical labels and continuous affect dimensions. In practice, an emotion category such as anger or sadness is not associated with a single fixed point in affective space; rather, it occupies a dispersed and often multi-modal region whose shape depends on subject variability, intensity, and context. This mismatch makes rigid one-to-one constraints between class labels and dimensional targets overly simplistic. Recent work has therefore explored more flexible mechanisms for bridging categorical and dimensional representations, including distribution matching, category-guided dimensional inference, and data-driven affect priors [5–7]. In our previous work, this issue was addressed through Gaussian Mixture Model (GMM)-based soft consistency, which models each emotion class as a distribution in V-A-D space rather than as a fixed centroid [4]. This probabilistic view is particularly suitable when the underlying affective structure is heterogeneous and cannot be faithfully represented by hard deterministic mappings [8–10].

A second challenge, which is more directly tied to real-world deployment, is the absence of calibrated trust. Conventional FER systems typically output point predictions only, without indicating how reliable those predictions are. This is problematic because facial expressions in practice are often ambiguous, weakly expressed, or visually degraded, and models can therefore be confidently wrong in precisely the cases where reliability matters most. In the broader machine learning literature, uncertainty quantification (UQ) has emerged as a principled way to address this issue. Bayesian deep learning methods can distinguish uncertainty arising from limited model knowledge from uncertainty inherent in the data itself, thereby providing a richer notion of predictive confidence [11]. Representative approaches include Monte Carlo Dropout [12], Bayes by Backprop [13], Variational Dropout with the local reparameterization trick [14], and Deep Ensembles [15]. More recently, deterministic distance-aware alternatives have also been proposed for reliable uncertainty estimation under distribution shift [16]. These developments strongly suggest that trust-aware affective systems should move beyond point prediction and explicitly model predictive uncertainty [17].

Despite this progress, existing FER-oriented uncertainty studies have concentrated predominantly on on annotation-side ambiguity and training-time reliability. Representative examples include latent label distribution mining [18], suppression of uncertain or mislabeled samples [19], relative uncertainty learning as a measure of sample difficulty [20], uncertainty-aware label distribution learning [21], and multi-task assisted correction of uncertain facial expressions [22]. Earlier Bayesian FER work also argued that probabilistic classifiers are more suitable than purely deterministic alternatives for handling ambiguous expressions [23,24]. Taken together, these studies have substantially advanced uncertainty-aware FER; however, their main focus remains label ambiguity, noisy supervision, or instance difficulty. By contrast, the question of how predictive uncertainty behaves when the input itself is progressively degraded remains insufficiently understood within this line of work. This distinction is crucial because in real applications, facial imagery may suffer from partial information loss, occlusion, blur, low resolution, or other departures from ideal acquisition conditions [25]. A separate line of FER research has addressed such degradations from a robustness perspective, for instance, through confidence-weighted local prediction [26] or multi-angle feature aggregation [27]. However, these methods treat confidence primarily as a task-specific weighting mechanism for robust classification rather than as Bayesian predictive uncertainty decomposable into epistemic and aleatoric components.

Motivated by this gap, this paper investigates uncertainty-aware multi-task FER from the perspective of predictive reliability under controlled input degradation. Building on the probabilistic category-dimension coupling provided by GMM-based soft consistency, this work develops BGDC for joint V-A-D prediction and uncertainty-aware inference. Unlike

prior studies that mainly treat uncertainty as a mechanism for handling label ambiguity or filtering unreliable training samples, the present work uses controlled input degradation as a diagnostic intervention to examine whether uncertainty estimates remain reliable and interpretable when facial evidence is progressively removed.

In particular, this study examines a commonly assumed but insufficiently validated expectation: epistemic uncertainty should increase when the input evidence supporting affect prediction becomes weaker. This expectation is evaluated through comparisons between BGDC and representative uncertainty estimation baselines under different spatial degradation settings. Two complementary forms of uncertainty are modeled: epistemic uncertainty, which reflects uncertainty in the learned predictive mapping, and aleatoric uncertainty, which reflects input-dependent observational noise. Their responses are analyzed jointly with prediction error, allowing uncertainty to be assessed not merely as an internal Bayesian quantity but also as a practical signal of predictive reliability.

The main contributions of this work are summarized as follows.

- A unified multi-task affective framework is established by coupling GMM-based soft consistency with Bayesian regression. Within this formulation, categorical emotion structure, continuous V-A-D prediction, and predictive uncertainty are handled in a single model rather than in separate stages.
- The framework is further extended to represent both epistemic and aleatoric uncertainty explicitly. As a result, affect prediction is no longer restricted to deterministic point estimates. The emphasis shifts toward reliability-aware inference, where the model is expected not only to predict affective states but also to indicate when its V-A-D predictions may be unreliable under ambiguous or non-ideal facial observations.
- A controlled degradation protocol is introduced as a diagnostic evaluation setting for uncertainty-aware FER. By progressively removing facial evidence under different spatial configurations, this protocol makes it possible to test whether epistemic uncertainty from different estimators remains aligned with prediction error and whether such responses depend on the location and task relevance of the removed facial region rather than only on the masking ratio. This reframes uncertainty as an interpretable indicator of predictive reliability under non-ideal facial observations rather than as an internal Bayesian quantity alone.

2. Related Work

2.1. Multi-Task Emotion Recognition with Categorical and Dimensional Supervision

Emotion recognition has traditionally been studied through two major representational paradigms: discrete categorical labels and continuous affective dimensions. Categorical models provide intuitive supervisory targets such as anger, happiness, and sadness, whereas dimensional models aim to describe emotion as a point or region in a continuous affective space. In modern affective computing, these two paradigms are increasingly viewed as complementary rather than competing, since category labels offer semantic interpretability while dimensional targets provide finer-grained variation within and across categories [1–3]. This view has motivated the rise of multi-task learning for emotion recognition, where both forms of annotation are exploited simultaneously. Recent studies have shown that categorical and dimensional emotion representations are statistically coupled. For example, category-guided dimensional emotion detection and circumplex-guided inference have demonstrated that category supervision can regularize the prediction of continuous affect and vice versa [6,7]. Distribution matching has further shown that the relationship among multiple emotion tasks can be encoded through structured alignment in representation space [5]. In the image emotion literature, continuous probability distribution prediction has also highlighted the value of representing affective targets as

distributions rather than fixed labels [9]. More recently, multi-dimensional emotion recognition based on category labels and VAD targets has provided explicit evidence that emotion categories can be constrained through their positions in continuous affective space, thereby improving joint prediction performance [3].

However, most such approaches still rely on relatively rigid assumptions about the relationship between discrete and continuous emotion structure. In practice, facial expressions belonging to the same category frequently form diffuse, overlapping, and multi-modal clusters in affective space. This issue becomes more pronounced when dominance is included in addition to valence and arousal because within-class variation cannot be well approximated by a single deterministic target. This limitation was previously addressed by introducing GMM-based soft consistency, whereby each class is represented by a learned distribution over V-A-D space rather than by a single point [4]. Compared with rule-based or point-based consistency constraints, this formulation offers a more flexible probabilistic bridge between categorical and dimensional supervision. It is also consistent with broader efforts to revisit emotion representation in the wild and to move toward richer, distribution-aware affective models [10].

2.2. Uncertainty Quantification in Deep Learning

Uncertainty quantification has become a central topic in modern deep learning because high predictive accuracy alone is insufficient for reliable deployment in open-world settings. This concern is especially relevant to affective computing, where model outputs may directly influence user interaction quality and system trust; accordingly, recent work has emphasized that affective systems should move beyond point-estimate prediction and provide interpretable and reliability-aware outputs [28]. The most widely adopted conceptual framework distinguishes between aleatoric uncertainty, which arises from noise or ambiguity intrinsic to the observation, and epistemic uncertainty, which arises from uncertainty in model parameters and can in principle be reduced by additional data [11]. This decomposition is particularly suitable for emotion recognition, where both ambiguous observations and limited training coverage are common.

A large body of work has proposed practical methods for estimating these uncertainties. MC Dropout interprets standard dropout training as approximate Bayesian inference and enables uncertainty estimation through repeated stochastic forward passes [12]. Bayes by Backprop explicitly learns posterior distributions over weights by optimizing a variational free-energy objective [13]. Variational Dropout further improves the efficiency of approximate Bayesian learning through the local reparameterization trick and learnable dropout rates [14]. Deep Ensembles depart from explicit Bayesian formalism, but remain a strong and competitive baseline by aggregating independently trained models [15]. Deterministic alternatives such as distance-aware deep learning methods have also been proposed to approximate predictive uncertainty without sampling-based inference [16]. A recent review summarizes these families within a broader taxonomy of uncertainty representations, including Bayesian, ensemble-based, evidential, and interval-based paradigms [17].

Within affective computing, probabilistic modeling has been explored as a way to capture ambiguity and improve predictive reliability [29,30]. Related reliability concerns also appear in broader human-centered perception, where complementary visual modalities and temporal context have been used to stabilize recognition under visually ambiguous inputs [31]. While that line of work approaches reliability from the perspective of multimodal activity understanding, the present study examines a complementary setting in which probabilistic uncertainty is explicitly modeled for facial emotion classification and V-A-D regression under controlled facial evidence degradation. This distinction motivates

the central question of this study: whether uncertainty signals remain meaningful when facial evidence becomes incomplete or degraded.

2.3. Predictive Reliability and Uncertainty in Facial Expression Recognition

Within FER, uncertainty has been approached from several directions, although most existing studies remain centered on label ambiguity or training-time reliability rather than on systematic degradation of the visual input. Early Bayesian FER work established the basic motivation. Matin and Valdenegro-Toro argued that Bayesian neural networks are better aligned with the soft and uncertain nature of human emotion annotation than deterministic classifiers [23]. Earlier probabilistic FER methods also employed Bayesian network classifiers to combine labeled and unlabeled data [24]. The broader implication is straightforward: FER need not always be treated as a hard-label recognition problem.

Later work moved this idea closer to concrete training strategies. DMUE mines latent label distributions and models pairwise uncertainty among samples, addressing ambiguity at the supervision level [18]. SCN suppresses uncertain or potentially mislabeled instances through ranking regularization and self-correction [19]. Relative Uncertainty Learning reframes reliability as comparative sample difficulty and offers a lightweight alternative to full probabilistic inference [20]. Uncertainty-aware label distribution learning reconstructs soft targets from auxiliary affective structure, so that annotator disagreement becomes part of the supervision itself [21]. Multi-task assisted correction follows a related direction by using auxiliary objectives to improve recognition on uncertain expressions [22]. More recent noisy-label FER work has further combined noisy-sample extraction, pseudo-label correction, and robust retraining to reduce the influence of corrupted supervision [32]. Taken together, these studies leave little doubt that uncertainty is central to FER. It is not an auxiliary concern.

A different line of work addresses partial facial information more directly, but from the perspective of robustness rather than probabilistic uncertainty. Confidence-weighted local prediction reduces the influence of unreliable facial regions and aggregates evidence from the remaining visible areas [26]. Multi-angle feature extraction has likewise been introduced to compensate for occluded frontal views by incorporating complementary spatial perspectives [27]. These approaches clearly show that occlusion is a practically important challenge for FER.

That distinction becomes important once facial evidence is progressively removed. At that point, the question is no longer whether the annotation itself is noisy, but whether the model's uncertainty output changes in a way that faithfully reflects the severity and spatial pattern of information loss. This matters under realistic operating conditions, where FER must often function under appearance variation, unconstrained capture conditions, low image resolution, and partial visibility of diagnostically important facial regions [2,25,33]. Recent low-resolution FER research further shows that degraded visual quality can induce data uncertainty in addition to label ambiguity, motivating probabilistic embeddings for ambiguous low-quality facial images [33]. A reliability-aware system should therefore treat uncertainty as an operational warning signal when the input moves away from the effective support of the training distribution. Existing FER uncertainty studies provide the conceptual basis for reliability-aware learning, while occlusion-robust FER methods establish the practical significance of partial information loss [18–22,26,27].

The remaining limitation is therefore not the absence of occlusion-robust FER methods or confidence-aware mechanisms, but the lack of systematic analysis of how Bayesian predictive uncertainty itself responds when facial evidence is progressively and spatially removed at inference time. The present work addresses this question by treating uncertainty as a measurable response variable under controlled facial information loss, linking multi-

task affective modeling, Bayesian uncertainty estimation, and predictive reliability analysis within a unified framework.

3. Methodology

3.1. Dual-Branch Architecture and Uncertainty-Aware Emotion Representation

Facial emotion recognition is formulated as a joint prediction problem over two complementary targets: discrete emotion categories and continuous V-A-D coordinates. Though both are inferred from the same facial image, they place different demands on the visual representation. Category recognition prioritizes discriminative semantic boundaries, whereas dimensional regression must preserve subtle affective variation within and across those boundaries. A single shared prediction stream is unlikely to serve both objectives well, and a dual-branch architecture is therefore adopted.

The overall framework is illustrated in Figure 1. The model comprises a classification backbone, a regression backbone, an aleatoric head, a context-conditioned Bayesian regression head, and a semantic consistency module defined over the continuous affective space. These components form a unified multi-task pipeline in which categorical prediction, dimensional estimation, and uncertainty quantification are learned jointly.

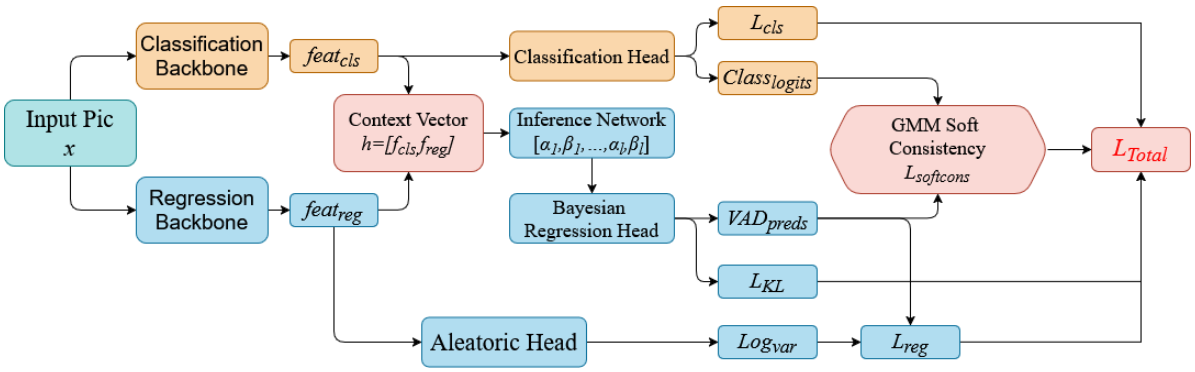


Figure 1. Multi-Task Emotion Recognition Network Framework.

Let the input facial image be $x \in \mathbb{R}^{H\times W\times C_{\text{in}}}$, where H, W, and C_{in} denote the height, width, and number of channels, respectively. The discrete emotion label is $c \in \{1,\ldots,C_e\}$, and the continuous affective target is

$$y = [y_{\text{V}}, y_{\text{A}}, y_{\text{D}}]^{\top} \in \mathbb{R}^3 \tag{1}$$

where y_{V}, y_{A}, and y_{D} are the Valence, Arousal, and Dominance annotations, respectively.

Two task-specific feature extractors are employed. The classification branch produces $f_{cls}(x) \in \mathbb{R}^d$ and the regression branch yields $f_{reg}(x) \in \mathbb{R}^d$, where d is the backbone output dimension. These two embeddings are concatenated into a unified context vector,

$$h = [f_{cls}(x),\ f_{reg}(x)] \in \mathbb{R}^{2d} \tag{2}$$

which summarizes both categorical and dimensional evidence and later conditions the Bayesian regression head.

Uncertainty in this framework is decomposed into two components. Aleatoric uncertainty characterizes ambiguity inherent in the observation itself, such as weak expression intensity, local corruption, annotation dispersion, or incomplete facial evidence. Epistemic uncertainty reflects instability in the model's regression mapping and becomes prominent when the available evidence does not stably support a particular affective output. Since these two sources arise from different mechanisms, they are modeled separately.

3.2. Aleatoric Uncertainty Modeling

Not all facial images carry equally reliable affective cues. As one component of reliability-aware inference, the model is therefore required to express observation-level ambiguity as part of its predictive output. Some samples present clear and coherent evidence, whereas others contain only partial or visually degraded information. To account for such variability, an input-dependent observation model is introduced.

Let $\mu(x) = [\mu_V(x), \mu_A(x), \mu_D(x)]^\top \in \mathbb{R}^3$ denote the predicted V-A-D mean for input x, where $\mu_V(x)$, $\mu_A(x)$, and $\mu_D(x)$ correspond to the predicted Valence, Arousal, and Dominance values, respectively. To characterize input-dependent observation noise in the predictive distribution, an aleatoric head $g_{ale}(\cdot)$ is attached to the regression branch. Taking the regression embedding $f_{reg}(x)$ as input, it produces a three-dimensional log-variance vector

$$s(x) = g_{ale}(f_{reg}(x)) = [s_V(x),\, s_A(x),\, s_D(x)]^\top \in \mathbb{R}^3 \tag{3}$$

whose three components correspond to the Valence, Arousal, and Dominance dimensions, respectively. More generally, for $j \in \{V, A, D\}$,

$$s_j(x) = \log \sigma_j^2(x) \tag{4}$$

where $\sigma_j^2(x) > 0$ denotes the predicted observation variance of the j-th affective dimension. Parameterizing uncertainty in terms of log-variance rather than variance ensures positivity of $\sigma_j^2(x)$ without requiring an explicit output constraint on $g_{ale}(\cdot)$.

Under the conditional-independence assumption across the three affective dimensions, the aleatoric observation covariance is defined as

$$\Sigma_a(x) = \mathrm{diag}\Big(\sigma_V^2(x),\, \sigma_A^2(x),\, \sigma_D^2(x)\Big) \in \mathbb{R}^{3\times 3} \tag{5}$$

where the subscript a indicates the aleatoric component of predictive uncertainty. The continuous affective target is then modeled as

$$p(y \mid x) = \mathcal{N}(y; \mu(x), \Sigma_a(x)) \tag{6}$$

The corresponding heteroscedastic negative log-likelihood (NLL) is adopted as the aleatoric regression loss. For a single sample (x, y), the loss is written as

$$\mathcal{L}_{ale}(x, y) = \frac{1}{2} \sum_{j=1}^{3} \Big[\exp(-s_j(x))(y_j - \mu_j(x))^2 + s_j(x)\Big] \tag{7}$$

where y_j, $\mu_j(x)$, and $s_j(x)$ denote the j-th components of y, $\mu(x)$, and $s(x)$, respectively. For a mini-batch $\mathcal{B} = \{(x_n, y_n)\}_{n=1}^N$, the batch-level loss is defined as

$$\mathcal{L}_{ale}^{\text{batch}} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{ale}(x_n, y_n) \tag{8}$$

This formulation admits a direct probabilistic interpretation. The squared residual of an ambiguous sample is down-weighted by its predicted variance, while the log-variance term penalizes the degenerate solution of assigning uniformly large uncertainty to all inputs. In practice, the predicted log-variance is clipped to a bounded interval to improve numerical stability during training.

Optimizing $\mathcal{L}_{ale}^{\text{batch}}$ alone can nevertheless admit a degenerate optimum in which the aleatoric head inflates $s(x)$ to down-weight every residual, rather than improving the regression mean $\mu(x)$. To avoid this, an auxiliary mean-squared-error term is introduced

to anchor $\mu(x)$ to the ground truth. The regression loss effectively used during training is therefore defined as

$$\mathcal{L}_{reg} = \mathcal{L}_{ale}^{\text{batch}} + \frac{1}{N}\sum_{n=1}^{N} \|\mu(x_n) - y_n\|_2^2 \tag{9}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. The second term anchors the regression mean to the ground truth, while the first preserves the probabilistic interpretation of aleatoric uncertainty.

3.3. Epistemic Uncertainty Modeling

Aleatoric uncertainty is attached to the observation, whereas epistemic uncertainty is associated with the model parameters. This distinction becomes particularly important when visible evidence is incomplete since the regression head may then fail to produce a stable affective mapping even when observation ambiguity is not dominant. For that reason, the regression head is formulated in Bayesian form.

The Bayesian regression head comprises L stacked Bayesian linear layers. For each layer $l \in \{1, \ldots, L\}$, an independent inference network produces a pair of scalar scales $(\alpha_l(h), \beta_l(h)) \in \mathbb{R}_+ \times \mathbb{R}_+$ that condition its parameter posterior on the global context h. To simplify notation, the layer index l is omitted in the equations below; Equations (10)–(23) describe a single representative Bayesian layer.

Consider a Bayesian linear layer inside the Bayesian regression head. Let $a \in \mathbb{R}^{d_{in}}$ denote its input activation, and let $z \in \mathbb{R}^{d_{out}}$ denote the corresponding layer output. The affine transformation is written as

$$z = aW + b \tag{10}$$

where $W \in \mathbb{R}^{d_{in} \times d_{out}}$ and $b \in \mathbb{R}^{d_{out}}$ are treated as random variables rather than fixed parameters. Here, z denotes the output of the Bayesian layer rather than the final V-A-D prediction.

Exact posterior inference over (W, b) is intractable, so a variational approximation $q_\phi(W, b \mid h)$ is introduced. An auxiliary inference network receives the global context vector h defined in Equation (2) and produces two nonnegative scalar scales,

$$\alpha(h),\ \beta(h) \in \mathbb{R}_+ \tag{11}$$

which jointly control the posterior spread of the weights and biases, respectively. The corresponding variational posteriors are defined as

$$q_\phi(W \mid h) = \mathcal{N}\Big(W;\ M,\ \text{diag}\Big(\alpha(h)^2 \odot M^2\Big)\Big) \tag{12}$$

$$q_\phi(b \mid h) = \mathcal{N}\Big(b;\ b_0,\ \text{diag}\Big(\beta(h)^2\Big)\Big) \tag{13}$$

where $M \in \mathbb{R}^{d_{in} \times d_{out}}$ and $b_0 \in \mathbb{R}^{d_{out}}$ are the variational means. Here, $\text{diag}(\cdot)$ denotes the diagonal covariance structure of the parameter posterior, and the scalar $\alpha(h)^2$ (resp. $\beta(h)^2$) is broadcast across the corresponding parameter block to modulate its variance uniformly.

A structured Gaussian prior is imposed on the parameters:

$$p(W) = \mathcal{N}\Big(W;\ 0,\ \text{diag}\Big(\sigma_p^2 \odot M^2\Big)\Big) \tag{14}$$

$$p(b) = \mathcal{N}(b;\ 0,\ \sigma_b^2 I) \tag{15}$$

where σ_p^2 and σ_b^2 are predefined scalar prior variance hyperparameters, and I denotes the identity matrix of appropriate size. Weights and biases are assumed independent, so the joint prior is written as

$$p(W, b) = p(W)p(b) \tag{16}$$

Learning proceeds by minimizing the negative evidence lower bound over the training set

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{17}$$

The variational objective is written as

$$\mathcal{L}_{VI} = -\mathbb{E}_{q_\phi(W,b|h)}[\log p(\mathcal{D} \mid W, b)] + \mathcal{L}_{KL} \tag{18}$$

where

$$\mathcal{L}_{KL} = \mathrm{KL}\big(q_\phi(W, b \mid h) \,\|\, p(W, b)\big) \tag{19}$$

denotes the variational regularization term imposed on the Bayesian regression branch. Using the regression loss introduced in Equation (9), the variational objective can be written in the compact form

$$\mathcal{L}_{VI} = \mathcal{L}_{reg} + \mathcal{L}_{KL} \tag{20}$$

which is the quantity optimized by the Bayesian regression branch during training.

To reduce gradient variance during training, local reparameterization is applied. Under the posterior in Equations (12) and (13), the Bayesian layer output follows a Gaussian distribution with

$$\mu_z = aM + b_0 \tag{21}$$

$$\sigma_z^2 = (a^2)\Big(\alpha(h)^2 \odot M^2\Big) + \beta(h)^2 \tag{22}$$

where a^2 denotes the element-wise square of the input activation vector. A stochastic sample is then obtained by

$$z = \mu_z + \sigma_z \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{23}$$

At inference time, the Bayesian regression head performs multiple stochastic forward passes for each input. The final V-A-D prediction is obtained by averaging the sampled outputs, while epistemic uncertainty is estimated from their dispersion. This uncertainty measure is used in the subsequent controlled masking and degradation analyses.

3.4. Semantic Consistency in the Continuous Affective Space

Accurate prediction and uncertainty estimation do not by themselves guarantee that the predicted V-A-D state remains consistent with the discrete emotion label. To preserve this category-conditioned structure, a class-conditional GMM is introduced in the V-A-D space. Unlike the aleatoric and epistemic components, this term does not quantify predictive uncertainty. It instead regularizes the semantic consistency of the continuous affective output.

For each emotion class c, the corresponding V-A-D density is modeled as

$$p(v \mid c; \Theta_c) = \sum_{k=1}^{K_c} \pi_{c,k}\, \mathcal{N}(v \mid \mu_{c,k}, \Sigma_{c,k}) \tag{24}$$

where $\Theta_c = \{\pi_{c,k}, \mu_{c,k}, \Sigma_{c,k}\}_{k=1}^{K_c}$ is the parameter set of the mixture. Rather than enforcing a hard class-conditioned penalty anchored to the ground-truth label, the semantic consistency constraint is coupled to the classifier output so that multiple plausible emotion categories can jointly regularize the V-A-D prediction. Let $\hat{y}_n$ denote the predicted V-A-D output

of the n-th sample and $p(c \mid x_n)$ the classifier's predicted posterior over the C_e emotion classes. The GMM-based soft semantic consistency loss is defined as

$$\mathcal{L}_{sem} = -\frac{1}{N}\sum_{n=1}^{N}\log\sum_{c=1}^{C_e} p(c \mid x_n)\, p(\hat{y}_n \mid c;\Theta_c) \tag{25}$$

Minimizing this term encourages the predicted continuous affective state to remain aligned with the emotion-conditioned affective structure, weighted by the model's own categorical belief rather than by the hard ground-truth label. This soft formulation provides a more forgiving training signal when the classification is ambiguous, while still concentrating the gradient on the class that the classifier considers most likely.

The class-conditional GMMs $\{\Theta_c\}_{c=1}^{C_e}$ are estimated once, prior to neural network training, by running expectation-maximization on the training-set (emotion label, V-A-D) pairs. Their parameters are then held fixed throughout the subsequent end-to-end optimization, so that $\mathcal{L}_{sem}$ serves as a stable category-conditioned prior rather than a jointly optimized component.

3.5. Joint Training Objective

The complete optimization objective integrates all four terms:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{kl}\mathcal{L}_{KL} \tag{26}$$

where λ_{cls}, λ_{reg}, λ_{sem}, and λ_{kl} are nonnegative weighting coefficients. The term $\mathcal{L}_{cls}$ denotes the discrete emotion classification loss, $\mathcal{L}_{reg}$ the uncertainty-aware regression loss, $\mathcal{L}_{sem}$ the GMM-based semantic consistency loss, and $\mathcal{L}_{KL}$ the variational regularization term. All model parameters, including both task-specific backbones and the Bayesian regression head, are optimized end-to-end using AdamW to minimize $\mathcal{L}_{total}$.

4. Experimental Validation

4.1. Experimental Environment and Dataset

All experiments were conducted on a workstation with an NVIDIA RTX 3090 GPU (24 GB VRAM), a 14-core Intel Xeon CPU, and 60 GB of RAM. The operating system was Ubuntu 22.04 with CUDA version 12.4. For model training, a pre-trained ResNet-18 served as the backbone network. The following hyperparameters were used: batch size of 32, initial learning rate of 0.005, and early stopping with a patience of 15 epochs.

The dataset used in this study is the publicly available FER2013 Facial Expression Recognition dataset, as augmented with continuous three-dimensional V-A-D annotations in the work of Huo and Ge [3]. The dataset comprises 28,709 training samples, 3589 validation samples, and an equal number of test samples. Each sample contains a discrete emotion class label and a three-dimensional continuous V-A-D annotation with values in the $[-2, 2]$ range. With the specified hardware and dataset, the training time for a single run was approximately 40 min.

Implementation Details

The regression backbone output dimension was $d = 512$, and the context vector used by the Bayesian regression head was formed by concatenating the classification and regression embeddings, resulting in $h \in \mathbb{R}^{1024}$. The context-conditioned Bayesian regression head contained two Bayesian linear layers, $512 \to 256 \to 3$. The aleatoric head was implemented as a lightweight multilayer perceptron that maps the 512-dimensional regression embedding to a three-dimensional log-variance vector for Valence, Arousal, and Dominance. During final inference, BGDC and MC Dropout used 100 stochastic forward passes

to estimate the predictive mean and epistemic uncertainty. The class-conditional GMM prior was fixed after fitting and was not updated during neural network training. Deep Ensemble consisted of five independently trained deterministic MSE baseline members, and its epistemic uncertainty was computed from the variance across the five member predictions. Unless otherwise specified, the final loss weights were set to $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{reg}} = 1.0$, $\lambda_{\text{sem}} = 0.005$, and $\lambda_{\text{MSE}} = 1.2$. The KL coefficient was linearly warmed up to $\lambda_{\text{kl}} = 0.015$ over the first five epochs to avoid imposing a strong variational regularization penalty before the V-A-D regression mean became stable.

4.2. Experimental Setup

4.2.1. Controlled Masking Settings

To examine whether uncertainty estimates remain reliable under controlled input degradation, a synthetic masking protocol is introduced on the FER2013-VAD test set. The purpose of this protocol is not merely to evaluate whether prediction accuracy decreases when facial images are partially corrupted. Rather, it is designed as a diagnostic setting for testing whether different uncertainty estimators produce epistemic responses that remain aligned with prediction error when local facial evidence is systematically removed.

Rather than modifying the semantic labels or the ground-truth V-A-D annotations, the input image itself is partially masked according to predefined geometric rules. This setting provides a controlled intervention on the available visual evidence while keeping the affective targets unchanged. It therefore makes it possible to separate two related but distinct questions: whether prediction error increases under degraded inputs, and whether the estimated uncertainty changes in a manner that is consistent with the resulting decline in predictive reliability.

Four masking modes are considered: upward, downward, center, and random masking. These modes are used not only to vary the amount of removed pixels but also to examine whether uncertainty responses depend on the spatial location of the removed evidence. This distinction is important because the masking ratio alone is not necessarily equivalent to the loss of affectively relevant facial information.

Let $x \in \mathbb{R}^{C \times H \times W}$ denote an input image, where C, H, and W represent the number of channels, image height, and image width, respectively. Let $r \in [0, 1]$ denote the masking ratio, and let m denote the masking mode, where

$$m \in \{\text{up}, \text{down}, \text{center}, \text{random}\} \tag{27}$$

For each masking configuration (r, m), a binary mask

$$M^{(r,m)} \in \{0, 1\}^{H \times W} \tag{28}$$

is generated according to a predefined geometric rule. The value $M_{i,j}^{(r,m)} = 1$ indicates that the spatial location (i, j) is masked, whereas $M_{i,j}^{(r,m)} = 0$ indicates that the original pixel value is retained.

The masked image $\tilde{x}$ is defined as

$$\tilde{x} = x \odot (1 - M^{(r,m)}) + v\, M^{(r,m)} \tag{29}$$

where $\odot$ denotes element-wise multiplication, and v denotes the fixed replacement value assigned to the masked region. In the present experiments, masking is applied in the transformed image tensor space, and the masked pixels are replaced by a constant value $v = 0$.

4.2.2. Up Masking

The up masking mode masks a horizontal region from the upper part of the image. Let

$$h_r = \lfloor rH \rfloor \tag{30}$$

denote the height of the masked band. The corresponding binary mask is defined as

$$M_{i,j}^{(r,\text{up})} = \begin{cases} 1, & 0 \le i < h_r \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

for $i = 0, 1, \ldots, H-1$ and $j = 0, 1, \ldots, W-1$.

Under this setting, the masked area occupies the upper h_r rows and spans the full image width. The total masked area is therefore $h_r W$, which approximately corresponds to the prescribed ratio rHW.

4.2.3. Down Masking

The down masking mode masks a horizontal region from the lower part of the image. Using the same band height $h_r = \lfloor rH \rfloor$, the binary mask is defined as

$$M_{i,j}^{(r,\text{down})} = \begin{cases} 1, & H - h_r \le i < H \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

This setting masks the bottom h_r rows of the image while preserving the remaining region. As in the up masking case, the masked area is approximately controlled by the ratio r.

4.2.4. Center Masking

The center masking mode masks a square region centered in the image. To make the masked area approximately equal to rHW, the side length of the square mask is defined as

$$s_r = \left\lfloor \sqrt{rHW} \right\rfloor \tag{33}$$

The top-left coordinate of the square mask is given by

$$i_0 = \left\lfloor \frac{H - s_r}{2} \right\rfloor \qquad j_0 = \left\lfloor \frac{W - s_r}{2} \right\rfloor \tag{34}$$

The corresponding binary mask is then defined as

$$M_{i,j}^{(r,\text{center})} = \begin{cases} 1, & i_0 \le i < i_0 + s_r \text{ and } j_0 \le j < j_0 + s_r \\ 0, & \text{otherwise} \end{cases} \tag{35}$$

This mode is designed to mask the central visual structure of the face image, thereby simulating a structured loss of facial evidence around the image center.

4.2.5. Random Masking

The random masking mode also employs a square mask with side length $s_r = \lfloor \sqrt{rHW} \rfloor$, but its location is randomly sampled rather than fixed at the image center. Let (i_r, j_r) denote the top-left coordinate of the square mask. Both coordinates are drawn independently from the discrete uniform distribution $\mathcal{U}\{\cdot\}$ over the set of valid positions:

$$i_r \sim \mathcal{U}\{0, \ldots, H - s_r\} \qquad j_r \sim \mathcal{U}\{0, \ldots, W - s_r\} \tag{36}$$

The corresponding binary mask is defined as

$$M_{i,j}^{(r,\text{random})} = \begin{cases} 1, & i_r \le i < i_r + s_r \text{ and } j_r \le j < j_r + s_r \\ 0, & \text{otherwise} \end{cases} \tag{37}$$

Unlike the previous three structured settings, this mode introduces spatial variability into the masking process. It is therefore used to examine whether the uncertainty response depends only on the masked area ratio or also on the specific spatial position of the removed visual information.

4.3. Analysis of Overall Experimental Results

A comparative evaluation was conducted on the FER2013-VAD dataset across five model configurations: MSE Baseline, Aleatoric Baseline, MC Dropout, BGDC, and Deep Ensemble. To reduce the dependence of the comparison on a single random initialization, Table 1 reports mean ± standard deviation over three independent seeds, namely 42, 30, and 100. The evaluation covers classification accuracy, calibration, V-A-D regression, uncertainty-error alignment, and model complexity. Classification performance was measured by accuracy, calibration by Expected Calibration Error (ECE), regression by mean V-A-D CCC and mean V-A-D MSE, and uncertainty reliability by the Spearman rank correlation between predicted uncertainty and per-sample regression error.

Table 1. Comparative performance of different model configurations on FER2013-VAD over three independent seeds. Results are reported as mean ± standard deviation. Best results are shown in bold. The symbols ↑ and ↓ indicate that higher and lower values are preferred, respectively.

Method	Acc ↑	ECE ↓	Avg CCC ↑	Avg MSE ↓	ρ_{ep} ↑	ρ_{al} ↑	Params (M)
MSE Baseline	0.6760 ± 0.0081	0.0953 ± 0.0512	0.6010 ± 0.0033	1.0917 ± 0.0277	N/A	N/A	23.14
Aleatoric Baseline	0.6811 ± 0.0037	0.0225 ± 0.0011	0.6031 ± 0.0011	1.0585 ± 0.0133	-0.0007 ± 0.0248	$\mathbf{0.4393 \pm 0.0185}$	23.21
MC Dropout [15]	0.6782 ± 0.0041	0.0288 ± 0.0060	0.5996 ± 0.0033	1.0515 ± 0.0313	0.0635 ± 0.0458	0.4386 ± 0.0133	23.21
BGDC (Ours)	$\mathbf{0.6838 \pm 0.0091}$	0.0343 ± 0.0070	$\mathbf{0.6039 \pm 0.0057}$	1.0612 ± 0.0251	0.2674 ± 0.0848	0.4320 ± 0.0077	23.35
Deep Ensemble [29] (Members = 5)	0.6786 ± 0.0081	$\mathbf{0.0146 \pm 0.0020}$	0.5930 ± 0.0059	$\mathbf{0.9775 \pm 0.0034}$	$\mathbf{0.2843 \pm 0.0458}$	N/A	115.72

Table 1 shows that BGDC obtains the highest mean classification accuracy among the compared configurations, reaching 0.6838 ± 0.0091. The improvement over MSE Baseline, Aleatoric Baseline, MC Dropout, and Deep Ensemble is numerically positive, although paired tests across the three seeds do not indicate statistical significance for accuracy. The classification result should therefore be interpreted conservatively: the proposed Bayesian and semantic-consistency components do not degrade categorical recognition, but the present multi-seed evidence does not support a claim of statistically significant accuracy improvement.

For V-A-D regression, BGDC achieves the highest average CCC, 0.6039 ± 0.0057, whereas Deep Ensemble obtains the lowest average MSE, 0.9775 ± 0.0034. This confirms that CCC and MSE emphasize different properties of regression quality. MSE penalizes point-wise numerical deviation, whereas CCC evaluates concordance between prediction and ground truth by jointly considering correlation, variance, and mean shift. The BGDC result indicates better preservation of the global affective structure in the V-A-D space, while the Deep Ensemble result indicates stronger point-wise error minimization. Since dimensional emotion recognition is concerned not only with residual magnitude but also with the agreement between predicted and annotated affective structure, CCC remains a central indicator for evaluating V-A-D prediction quality.

The calibration results show a different pattern. Deep Ensemble achieves the lowest ECE, 0.0146 ± 0.0020, followed by Aleatoric Baseline (0.0225 ± 0.0011) and MC Dropout

(0.0288 ± 0.0060). BGDC obtains an ECE of 0.0343 ± 0.0070, which is substantially lower than the deterministic MSE Baseline but higher than the three other uncertainty-aware configurations. This indicates that the proposed model improves calibration over a purely deterministic baseline, but does not surpass the alternative uncertainty-aware estimators in classification calibration. This result is consistent with the expectation that Deep Ensemble, which aggregates predictions from five independently trained models, provides strong empirical calibration. It also suggests that, in this setting, explicit uncertainty modeling in the auxiliary branches is associated with better-calibrated classification outputs than the deterministic baseline.

The uncertainty-error alignment results provide the clearest evidence for the proposed Bayesian formulation. BGDC obtains an epistemic Spearman correlation of 0.2674 ± 0.0848, which is much higher than MC Dropout (0.0635 ± 0.0458) and Aleatoric Baseline (-0.0007 ± 0.0248). Paired comparisons across seeds show that the improvement of BGDC over MC Dropout in ρ_{ep} is statistically significant, with $\Delta = +0.2039$ and $p = 0.0188$. The improvement over Aleatoric Baseline is also significant, with $\Delta = +0.2681$ and $p = 0.0221$. These results indicate that the epistemic uncertainty produced by BGDC is more closely aligned with sample-wise regression error than the uncertainty signal obtained from MC Dropout or from a model without a dedicated epistemic mechanism.

The aleatoric Spearman correlation shows a complementary pattern. Aleatoric Baseline obtains the highest ρ_{al}, 0.4393 ± 0.0185, marginally above MC Dropout (0.4386 ± 0.0133) and BGDC (0.4320 ± 0.0077). Paired comparisons across the three seeds did not detect statistically significant differences between BGDC and the two baselines ($p = 0.6047$ against Aleatoric Baseline and $p = 0.6381$ against MC Dropout). The slightly lower aleatoric correlation in BGDC is therefore not strong evidence of a degraded aleatoric channel. Rather, it is consistent with the intended uncertainty decomposition of the model. Because BGDC additionally exposes an epistemic branch, part of the reliability-related variation can be represented through epistemic uncertainty, so the aleatoric component is less tightly coupled to aggregate regression residuals. Under a two-component uncertainty model, reliability information is expected to be distributed across complementary uncertainty sources rather than concentrated in a single channel.

Deep Ensemble achieves a slightly higher epistemic Spearman correlation than BGDC, 0.2843 ± 0.0458 versus 0.2674 ± 0.0848, but this comparison must be interpreted together with model complexity. Each Deep Ensemble member has approximately 23.14 M parameters, and the five-member ensemble therefore requires approximately 115.72 M parameters in total. In contrast, BGDC uses 23.35 M parameters, which is close to the single-model baselines. Thus, although Deep Ensemble remains the strongest method in terms of MSE, ECE, and marginally higher epistemic ranking, BGDC provides competitive uncertainty-error alignment within a single-model Bayesian formulation.

Overall, the multi-seed evaluation characterizes BGDC as a single-model reliability-aware alternative rather than as a uniformly dominant predictor. BGDC achieves the highest mean CCC and the highest mean classification accuracy among the compared configurations, and provides a substantially stronger epistemic uncertainty-error correspondence than MC Dropout. Deep Ensemble remains stronger in MSE and ECE and has a slightly higher epistemic ranking, but it requires five independently trained members. This performance profile is consistent with the goal of reliability-aware multi-task FER, where V-A-D concordance, uncertainty-error alignment, and computational cost are considered jointly rather than optimized as isolated metrics.

4.4. Cross-Dataset Transfer Validation on FERPlus

A cross-dataset transfer evaluation was conducted on FERPlus [34] to examine whether the framework remains informative under a related but distinct facial expression distribution. Since FERPlus does not provide three-dimensional V-A-D annotations, this experiment is not a cross-dataset V-A-D regression benchmark. It serves as a diagnostic transfer setting for categorical recognition, classification calibration, and uncertainty-based error detection.

Table 2 shows that Deep Ensemble obtains the highest classification accuracy on FERPlus, reaching 0.6729, followed by MC Dropout and Aleatoric Baseline. BGDC does not achieve the strongest classification transfer accuracy, indicating that the proposed framework should not be interpreted as a general-purpose cross-dataset classifier. In terms of classification calibration, MC Dropout achieves the lowest ECE, while BGDC does not outperform the alternative uncertainty-aware baselines.

Table 2. Cross-dataset validation on FERPlus. The evaluation focuses on classification accuracy, calibration, and epistemic uncertainty-based error detection. The symbols ↑ and ↓ indicate that higher and lower values are preferred, respectively.

Method	Acc ↑	ECE ↓	AUROC (Error, Epistemic) ↑	Spearman (Error, Epistemic) ↑	$AURC_{ep}$ ↓
MSE Baseline	0.6431	0.1103	N/A	N/A	N/A
Aleatoric Baseline	0.6502	0.0354	N/A	N/A	N/A
MC Dropout [15]	0.6539	0.0168	0.4250	−0.1236	0.3398
BGDC (Ours)	0.6331	0.0863	0.7657	0.4436	0.1774
Deep Ensemble [29] (Members = 5)	0.6729	0.0513	0.6492	0.2426	0.2204

The uncertainty-based error detection results show a different pattern. BGDC obtains the highest epistemic error-detection AUROC, 0.7657, the highest Spearman correlation between classification error and epistemic uncertainty, 0.4436, and the lowest epistemic AURC, 0.1774. These results indicate that, even when classification accuracy decreases under transfer, the epistemic branch of BGDC remains informative for identifying unreliable predictions. The FERPlus experiment therefore provides a limited but useful external validation: the main transferable property of BGDC is not superior classification accuracy, but the preservation of an epistemic uncertainty signal aligned with prediction failure.

4.5. Ablation and Sensitivity Analysis

4.5.1. Ablation of the Bayesian Posterior Parameterization

The Bayesian regression head can be parameterized at three increasingly expressive levels, and the role of each level was isolated through a controlled ablation. The three variants share the same backbone, classification head, GMM prior, aleatoric branch, training schedule, and default loss weights. They differ only in how the posterior scale parameters (α, β) are produced. In the fixed-variance variant (V2), α and β are fixed to predefined constants, so the posterior scale is neither learned nor conditioned on the input. This variant removes the adaptive posterior-scale mechanism while keeping the same Bayesian regression structure. In the global-scalar variant (V3), α and β are replaced by two learnable scalars shared across all inputs, allowing the posterior scale to be optimized but not to depend on the input. In the context-conditioned variant (V4), used in the full BGDC framework, $\alpha(h)$ and $\beta(h)$ are produced by the inference network conditioned on the context vector. The progression V2 → V3 → V4 therefore isolates two design choices: whether the posterior scale is learnable, and whether it is allowed to depend on the input.

The model complexity of the three variants is compared in Table 3. The context-conditioned formulation introduces only a marginal increase in parameters, from approximately 23.21 M to 23.35 M, while leaving the MAC count essentially unchanged.

The capacity gap between V2/V3 and V4 is therefore small and does not by itself explain the functional difference between the variants. What changes from V3 to V4 is not the size of the regression head, but whether the posterior scale is allowed to vary with the input.

Table 3. Model complexity of Bayesian posterior variants.

Variant	Params (M)	Trainable (M)	MACs (G)
V2 (Fixed-Variance)	23.21	23.21	3.648
V3 (Global Scalar)	23.21	23.21	3.648
V4 (Context-Conditioned BGDC)	23.35	23.35	3.648

To assess the predictive and uncertainty-related consequences of these parameterization choices, the three variants were trained from scratch with three independent seeds, namely 42, 43, and 44, under identical default loss weights. Classification accuracy, calibration, and V-A-D regression performance are summarized in Table 4, while uncertainty-error alignment, predictive negative log-likelihood, and risk-coverage behavior are summarized in Table 5.

Table 4. Predictive performance of Bayesian posterior variants on FER2013-VAD over three independent seeds. Results are reported as mean ± standard deviation. Variant definitions: V2 = Fixed-Variance, V3 = Global Scalar, and V4 = Context-Conditioned BGDC. The symbols ↑ and ↓ indicate that higher and lower values are preferred, respectively.

Variant	Acc ↑	ECE ↓	Avg CCC ↑	Avg MSE ↓
V2	0.6898 ± 0.0043	0.0337 ± 0.0022	0.5984 ± 0.0048	1.1067 ± 0.0068
V3	0.6840 ± 0.0036	0.0416 ± 0.0086	0.5953 ± 0.0038	1.1196 ± 0.0140
V4	0.6815 ± 0.0123	0.0314 ± 0.0099	0.6018 ± 0.0020	1.0918 ± 0.0497

Table 5. Uncertainty and risk-coverage behavior of Bayesian posterior variants on FER2013-VAD over three independent seeds. Results are reported as mean ± standard deviation. Variant definitions: V2 = Fixed-Variance, V3 = Global Scalar, and V4 = Context-Conditioned BGDC. The symbols ↑ and ↓ indicate that higher and lower values are preferred, respectively.

Variant	ρ_{ep} ↑	ρ_{al} ↑	NLL ↓	AURC ↓
V2	0.0893 ± 0.0130	0.4285 ± 0.0148	1.4127 ± 0.0080	0.6871 ± 0.0202
V3	0.0871 ± 0.0150	0.4237 ± 0.0151	1.4177 ± 0.0138	0.7065 ± 0.0207
V4	0.2699 ± 0.0340	0.4187 ± 0.0329	1.6099 ± 0.2344	0.7084 ± 0.0684

The results separate the contribution of posterior learning from the contribution of input conditioning. The transition from V2 to V3 introduces learnable posterior scales but keeps them input-independent, and no metric changes substantially between the two variants. The epistemic Spearman correlation remains almost unchanged, with 0.0893 ± 0.0130 for V2 and 0.0871 ± 0.0150 for V3. A paired t-test across the three seeds also does not detect a significant difference between them ($p = 0.786$). Making the posterior scale learnable but global is therefore not sufficient to obtain an uncertainty signal that tracks per-sample regression error.

The transition from V3 to V4 introduces input conditioning while keeping the rest of the architecture unchanged, and the epistemic uncertainty-error alignment changes substantially. The epistemic Spearman correlation increases to 0.2699 ± 0.0340, which is approximately three times larger than the values obtained by V2 and V3. Paired tests across the three seeds confirm that the improvement is statistically significant against both simpler variants, with $\Delta = +0.181$ and $p = 0.005$ against V2, and $\Delta = +0.183$ and $p = 0.010$ against V3. The uncertainty-error alignment reported in Section 4.3 is

therefore not produced by adding a Bayesian posterior in a generic sense, but specifically by conditioning the posterior scale on the input through the context vector.

Beyond ρ_{ep}, V4 also obtains the best mean ECE, Avg CCC, and Avg MSE among the three variants, as shown in Table 4. These differences are smaller than the improvement in ρ_{ep}, and they should therefore be interpreted as secondary evidence rather than as the main contribution of the context-conditioned posterior. V2 obtains the highest mean classification accuracy, but the gain in epistemic reliability from V4 does not produce a statistically significant loss in classification accuracy.

A trade-off appears in predictive negative log-likelihood. V2 obtains the lowest mean NLL (1.4127 ± 0.0080), V3 is almost identical to V2 (1.4177 ± 0.0138), and V4 is higher and more seed-variable (1.6099 ± 0.2344), as reported in Table 5. Paired tests do not detect a statistically significant NLL difference between V4 and the two simpler variants ($p = 0.280$ against V2 and $p = 0.283$ against V3). The seed-level pattern also shows that the V4 NLL increase is not uniform: the V4 run with seed 42 obtains NLL $= 1.355$, which is lower than every V2 and V3 run, whereas the V4 run with seed 44 obtains NLL $= 1.816$. Thus, the higher V4 mean NLL is mainly associated with larger seed-level variability.

This NLL behavior should be interpreted together with classification calibration and uncertainty ranking. V4 obtains the lowest mean ECE among the three variants, indicating that its higher NLL should not be read as direct evidence of worse classification calibration. A more conservative interpretation is a sharpness–reliability trade-off associated with input-adaptive posterior scaling. Predictive NLL rewards both calibration and sharpness, whereas the present reliability objective emphasizes whether uncertainty can rank samples according to prediction error. From this perspective, V4 sacrifices some likelihood sharpness and seed stability but gains a substantially stronger epistemic uncertainty-error correspondence.

Risk-coverage analysis provides a further qualification. As shown in Figure 2, V4 does not uniformly dominate V2 and V3 in risk-coverage behavior. The three curves are close over most coverage levels, and the AURC values in Table 5 slightly favor V2 on average. The main benefit of context conditioning is therefore not uniformly lower risk at all coverage levels but stronger sample-wise epistemic error ranking.

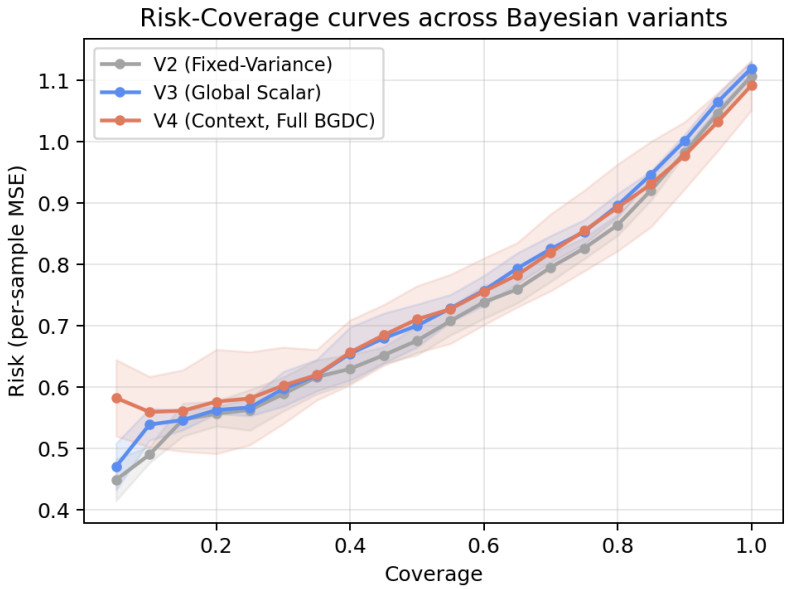


Figure 2. Risk-coverage curves of Bayesian posterior variants. The context-conditioned variant does not uniformly dominate the simpler variants in AURC; its main advantage lies in stronger epistemic uncertainty-error ranking rather than uniformly lower risk across all coverage levels. Shaded bands denote the ± 1 standard deviation range across independent seeds.

Taken together, these ablation results clarify the role of each design element in the Bayesian regression head. Posterior learning in isolation, represented by V3 versus V2, has no measurable effect on the reliability metrics. Input-dependent posterior parameterization, represented by V4 versus V3, is responsible for the substantial improvement in epistemic uncertainty-error correspondence, while model complexity is kept nearly unchanged. This benefit should be interpreted as an improvement in reliability-oriented uncertainty ranking, not as uniform dominance across all predictive, likelihood-based, and risk-coverage metrics.

4.5.2. Sensitivity to Loss Weights

A one-at-a-time sensitivity analysis was conducted for the semantic consistency weight λ_{sem}, the KL weight λ_{kl}, and the auxiliary MSE anchor λ_{MSE}. Unless the parameter under analysis was varied, the default setting was kept fixed at $\lambda_{\text{sem}} = 0.005$, $\lambda_{\text{kl}} = 0.015$, and $\lambda_{\text{MSE}} = 1.2$. Table 6 summarizes the results.

Table 6. One-at-a-time sensitivity analysis of the main loss weights. Default values are $\lambda_{\text{sem}} = 0.005$, $\lambda_{\text{kl}} = 0.015$, and $\lambda_{\text{MSE}} = 1.2$. The symbols ↑ and ↓ indicate that higher and lower values are preferred, respectively.

Family	Value	Acc ↑	ECE ↓	Avg CCC ↑	Avg MSE ↓	ρ_{ep} ↑	AURC ↓
λ_{sem}	0	0.6715	0.0221	0.6065	1.0327	0.2604	0.6426
λ_{sem}	0.001	0.6826	0.0210	0.6012	1.0305	0.2913	0.6412
λ_{sem}	0.005	0.6943	0.0395	0.6079	1.0661	0.2732	0.7156
λ_{sem}	0.025	0.6737	0.0334	0.5983	1.0239	0.2283	0.6541
λ_{sem}	0.1	0.6938	0.0213	0.5600	1.0415	0.2593	0.6385
λ_{kl}	0	0.6960	0.0234	0.6044	1.0879	0.2282	0.7248
λ_{kl}	0.003	0.6890	0.0433	0.6036	1.0617	0.2423	0.6984
λ_{kl}	0.015	0.6943	0.0395	0.6079	1.0661	0.2732	0.7156
λ_{kl}	0.075	0.6963	0.0259	0.5999	1.0871	0.2493	0.7299
λ_{kl}	0.5	0.6882	0.0311	0.5941	1.0764	0.3514	0.7076
λ_{MSE}	0	0.6882	0.0398	0.6052	1.0338	0.4269	0.6501
λ_{MSE}	0.3	0.6824	0.0272	0.6061	1.1066	0.3311	0.7222
λ_{MSE}	1.2	0.6943	0.0395	0.6079	1.0661	0.2732	0.7156
λ_{MSE}	3.0	0.6721	0.0304	0.6058	1.0710	0.0820	0.6660

The model is not invariant to the loss weights. Larger λ_{sem} values can preserve classification accuracy but may reduce V-A-D concordance, as observed at $\lambda_{\text{sem}} = 0.1$. The KL term affects the trade-off between posterior regularization and predictive accuracy: a very large KL weight increases ρ_{ep} but lowers CCC. The auxiliary MSE anchor also changes the uncertainty profile. Removing it increases ρ_{ep}, but the final setting $\lambda_{\text{MSE}} = 1.2$ yields the best combination of classification accuracy and mean CCC. The default configuration should therefore be interpreted as a balanced operating point rather than the optimum for every individual metric.

4.5.3. Task-Balance Ratio Between Classification and Regression

A separate task-balance sweep was conducted by varying the ratio $\lambda_{\text{reg}}/\lambda_{\text{cls}}$ while keeping λ_{sem}, λ_{kl}, and λ_{MSE} at their default values. Table 7 summarizes the results.

Table 7. Sensitivity to the task-balance ratio $\lambda_{\text{reg}}/\lambda_{\text{cls}}$ with other loss weights fixed at their default values. The symbols ↑ and ↓ indicate that higher and lower values are preferred, respectively.

$\lambda_{\text{reg}}/\lambda_{\text{cls}}$	Acc ↑	ECE ↓	Avg CCC ↑	Avg MSE ↓	NLL ↓	AURC ↓	ρ_{ep} ↑
0.5	0.6824	0.0143	0.6010	1.0009	1.3399	0.6216	0.3048
1.0	0.6943	0.0395	0.6079	1.0665	1.5102	0.7147	0.2739
2.0	0.6871	0.0312	0.6065	1.0736	1.5590	0.7198	0.2276

The task-balance sweep shows a trade-off between predictive performance and uncertainty-oriented reliability metrics. A smaller regression weight improves ECE, NLL, AURC, and ρ_{ep}, whereas the default ratio $\lambda_{reg}/\lambda_{cls} = 1.0$ yields the highest classification accuracy and mean CCC. Increasing the regression weight to 2.0 does not improve V-A-D concordance and further decreases ρ_{ep}. The default setting was therefore retained as a balanced operating point between categorical prediction and dimensional affective concordance.

4.6. Uncertainty Behavior Under Controlled Masking

After the overall predictive behavior and sensitivity of each model had been examined, attention was turned to whether uncertainty estimates provide reliable signals when facial information is progressively removed. Four masking modes, center, down, random, and up, were applied using a zero-filled mask with ratios $r \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. At each masking ratio, the models were evaluated using epistemic uncertainty, aleatoric uncertainty, and mean V-A-D MSE.

4.6.1. Aggregate Trends of Uncertainty Under Controlled Masking

In Figures 3–6, shaded bands indicate the $\pm 1\sigma$ range across samples, and ρ denotes the Spearman rank correlation between the plotted quantity and the masking ratio. Table 8 summarizes the ratio-level monotonicity patterns across all model-masking combinations. These coefficients differ from the sample-level Spearman values in Table 1: here they measure whether group-averaged uncertainty or error changes monotonically as the masking ratio increases.

Table 8. Ratio-level monotonicity of prediction error and uncertainty under controlled masking on FER2013-VAD. Entries are Spearman correlations ρ (masking ratio, mean metric) computed over seven masking ratios $\{0, 5, 10, 20, 30, 40, 50\}\%$. Best ρ_{ep} values per masking mode are shown in bold.

Method	Center Masking			Up Masking			Down Masking			Random Masking		
	ρ_{ep}	ρ_{al}	ρ_{mse}	ρ_{ep}	ρ_{al}	ρ_{mse}	ρ_{ep}	ρ_{al}	ρ_{mse}	ρ_{ep}	ρ_{al}	ρ_{mse}
Baseline	N/A	N/A	1.00	N/A	N/A	1.00	N/A	N/A	1.00	N/A	N/A	1.00
Aleatoric Baseline	N/A	0.93	1.00	N/A	−0.36	0.89	N/A	0.64	0.96	N/A	0.64	1.00
MC Dropout [15]	−0.96	1.00	1.00	−0.79	−0.21	0.96	−1.00	0.64	0.96	−1.00	1.00	1.00
BGDC (Ours)	0.89	0.57	1.00	−0.86	−0.93	0.96	1.00	0.54	0.96	1.00	0.14	1.00
Deep Ensemble [29]	−0.36	N/A	0.96	−0.14	N/A	1.00	−0.32	N/A	1.00	−0.46	N/A	1.00

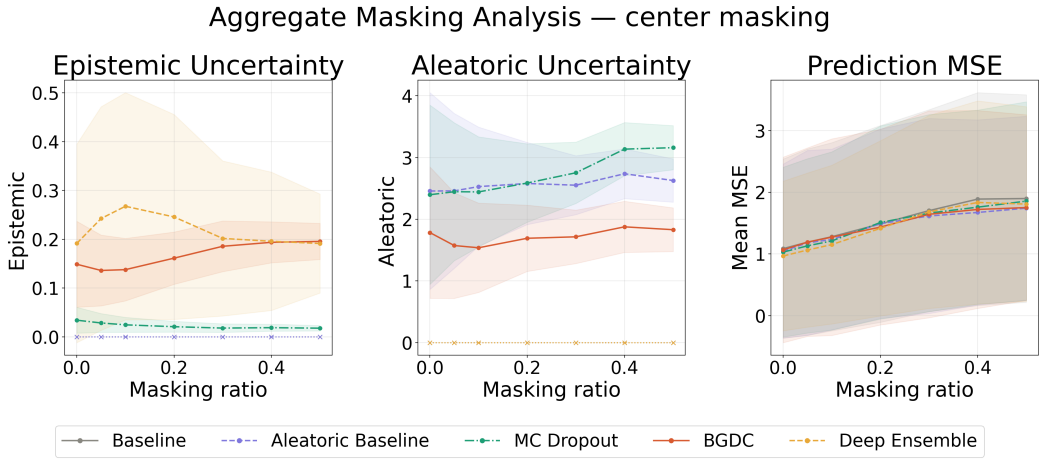


Figure 3. Uncertainty and MSE curves under center masking.

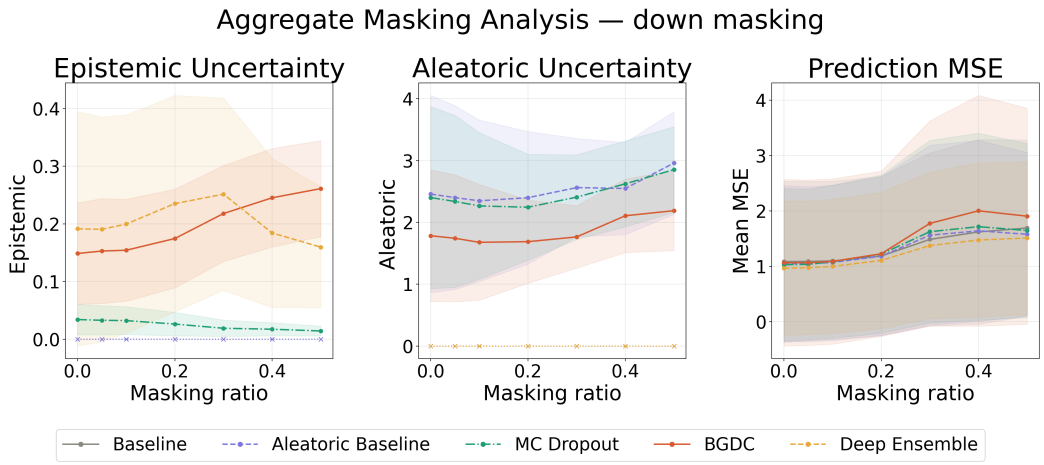


Figure 4. Uncertainty and MSE curves under downward masking.

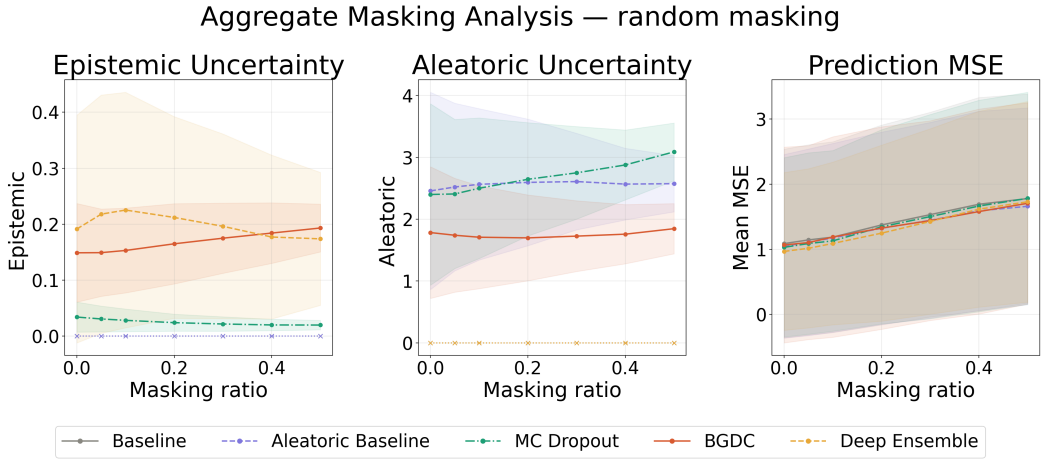


Figure 5. Uncertainty and MSE curves under randomly placed masking.

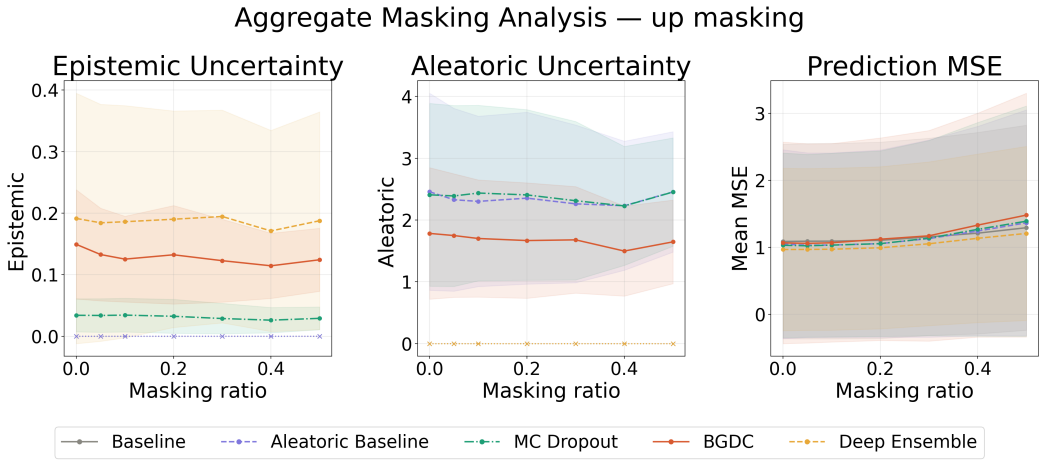


Figure 6. Uncertainty and MSE curves under upward masking.

Table 8 shows that prediction error generally increases under masking, as indicated by the consistently positive ρ_{mse} values. However, epistemic uncertainty does not increase monotonically for all estimators. MC Dropout shows negative epistemic monotonicity under all four masking modes, and Deep Ensemble also shows weak or negative ratio-level epistemic trends. In contrast, BGDC shows positive epistemic monotonicity under center, downward, and random masking, with $\rho_{\text{ep}} = 0.89$, 1.00, and 1.00, respectively. The exception is upward masking, where both epistemic and aleatoric uncertainty decrease. This exception is examined further below using emotion-specific evidence.

Figures 3–6 also show that the four masking modes are not equally destructive. Center and downward masking produce the largest MSE rise, random masking produces an intermediate response, and upward masking is the only mode where group-averaged epistemic uncertainty contracts despite a clear MSE increase. The mechanisms behind this region-dependent response are examined in Section 4.6.2.

The negative epistemic monotonicity of MC Dropout under structured masking can be explained by the interaction between dropout sampling and zero-filled spatial degradation. When facial regions are replaced by a zero-valued mask, the corresponding feature activations are suppressed toward zero. As the masked area grows, stochastic forward passes can therefore produce more similar outputs, causing dropout-induced predictive variance to contract even when the prediction task becomes harder. This indicates that dropout variance should not be treated as a default epistemic proxy under structured facial occlusion without validation in the target degradation regime.

4.6.2. Representative Case Study and Region-Dependent Evidence

Dataset-level masking curves describe the average response over the test set, but they may obscure how individual facial images are affected by different spatial perturbations. To provide a concrete visual reference, a representative Happy-class test instance is first examined under the four masking configurations. This visualization is used only to illustrate the spatial meaning of the masks; the quantitative support for the interpretation is provided by the compact emotion-level summary in Table 9 and by the context-vector and saliency-overlap diagnosis described below.

Figure 7 shows that the selected image contains an infant face with a clear smile, a wool hat occupying much of the upper region, and strong lower-face cues around the mouth and chin. Center masking progressively affects the central facial configuration, while downward masking directly interferes with the mouth and lower cheek regions that support the Happy interpretation. Upward masking initially covers the hat and upper forehead and therefore removes visible pixels without immediately damaging the most decisive smile-related evidence. Random masking is less spatially consistent because the masked square may overlap either expression-relevant regions or less informative image areas. This example illustrates why the same masking ratio does not necessarily correspond to the same amount of affective evidence loss.

An emotion-specific analysis was performed over representative upper-face-sensitive and lower-face-sensitive categories so that the region-dependent response is verified beyond the single Happy sample above. Fear and Surprise were treated as upper-face-sensitive categories, and Happy and Disgust as lower-face-sensitive categories. Table 9 reports the ratio-level Spearman correlation between masking ratio and epistemic uncertainty for these category groups.

Table 9 shows that the observed masking response is not driven by the selected Happy sample alone. For BGDC, downward masking produces strongly positive epistemic monotonicity for both upper-face and lower-face groups, with the lower-face group reaching a mean value of 1.000. Center and random masking also yield stronger positive responses for lower-face-sensitive categories. In contrast, upward masking produces negative epistemic monotonicity for Fear and Surprise, while the lower-face categories remain non-negative on average. This pattern indicates that upward masking does not behave as a uniformly reliable degradation condition since pixel removal, prediction error, and posterior spread can become partially decoupled either when the removed region is weakly diagnostic for the current expression or when the context-conditioned posterior contracts despite the removal of task-relevant upper-face evidence.

Table 9. Compact per-emotion summary of epistemic monotonicity under controlled masking. Values are ratio-level Spearman correlations between masking ratio and epistemic uncertainty. All uncertainty-aware models are summarized under the four masking modes to verify that the observed region-dependent response is not determined by a single selected sample.

Model	Masking Mode	Upper-Face Mean	Lower-Face Mean	Representative Upper Classes	Representative Lower Classes
BGDC	Center	0.375	0.946	Fear: 0.143, Surprise: 0.607	Happy: 0.929, Disgust: 0.964
BGDC	Up	−0.875	0.286	Fear: −0.857, Surprise: −0.893	Happy: 0.000, Disgust: 0.571
BGDC	Down	0.946	1.000	Fear: 0.893, Surprise: 1.000	Happy: 1.000, Disgust: 1.000
BGDC	Random	0.500	1.000	Fear: 0.107, Surprise: 0.893	Happy: 1.000, Disgust: 1.000
MC Dropout	Center	−0.893	−0.929	Fear: −0.964, Surprise: −0.821	Happy: −0.893, Disgust: −0.964
MC Dropout	Up	−0.875	−0.393	Fear: −0.893, Surprise: −0.857	Happy: −0.071, Disgust: −0.714
MC Dropout	Down	−0.982	−0.946	Fear: −0.964, Surprise: −1.000	Happy: −0.893, Disgust: −1.000
MC Dropout	Random	−0.982	−1.000	Fear: −1.000, Surprise: −0.964	Happy: −1.000, Disgust: −1.000
Deep Ensemble	Center	−0.429	−0.125	Fear: −0.750, Surprise: −0.107	Happy: 0.143, Disgust: −0.393
Deep Ensemble	Up	−0.214	−0.018	Fear: −0.393, Surprise: −0.036	Happy: −0.143, Disgust: 0.107
Deep Ensemble	Down	−0.321	0.089	Fear: −0.321, Surprise: −0.321	Happy: 0.500, Disgust: −0.321
Deep Ensemble	Random	−0.821	−0.214	Fear: −0.893, Surprise: −0.750	Happy: 0.214, Disgust: −0.643

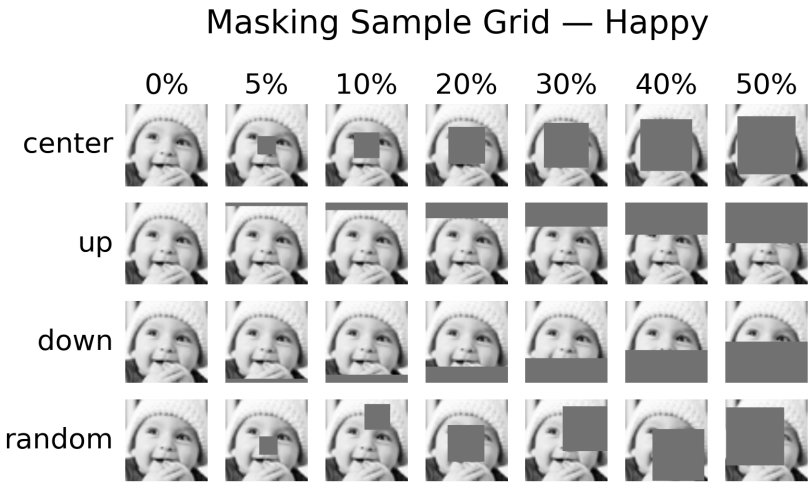


Figure 7. Visual progression of the four masking modes on the selected Happy sample. Columns correspond to masking ratios from 0% to 50%; rows correspond to mask placement.

This interpretation is further supported by the saliency-overlap diagnosis in Table 10. The diagnostic analysis examined whether changes in epistemic uncertainty were associated with the amount of diagnostically relevant facial evidence removed. For center, downward, and random masking, saliency overlap generally shows positive associations with ΔEp and ΔMSE, indicating that removing regions aligned with model-attended facial evidence tends to increase both prediction error and epistemic uncertainty. By contrast, upward masking produces a clear dissociation for Fear and Surprise: the overlap–ΔMSE correlations are strongly positive, 0.974 and 0.992, whereas the corresponding overlap–ΔEp correlations are negative, −0.880 and −0.723. This indicates that upper-region evidence removal can increase V-A-D prediction error while contracting the context-conditioned posterior spread. Therefore, the upward-masking case is treated as a boundary condition of epistemic reliability rather than as evidence of improved reliability.

Overall, the per-sample visualization, compact emotion-level summary, and saliency-overlap diagnosis support a bounded interpretation: BGDC provides meaningful reliability information in most structured degradation settings where the removed region overlaps diagnostically relevant facial evidence, while upward-masking posterior contraction must be read together with the spatial role of the removed region and the resulting context-conditioned posterior behavior, rather than as a standalone reliability claim. Masking grids

for Happy, Sad, and Surprise, together with the full per-sample matrix visualization, are provided in Appendixes A and B.

Table 10. Saliency-overlap diagnosis under controlled masking. Values are Pearson correlations between Grad-CAM saliency-overlap ratio and the changes in epistemic uncertainty or prediction error. ΔEp denotes the change in epistemic uncertainty, and ΔMSE denotes the change in V-A-D prediction error.

Masking Mode	Emotion	Overlap–ΔEp	Overlap–ΔMSE
Center	Disgust	0.944	0.858
Center	Fear	0.394	0.953
Center	Happy	0.969	0.970
Center	Surprise	0.662	0.824
Down	Disgust	0.976	0.891
Down	Fear	0.975	−0.311
Down	Happy	0.968	0.940
Down	Surprise	0.998	0.681
Random	Disgust	0.946	0.882
Random	Fear	0.237	0.969
Random	Happy	0.996	0.994
Random	Surprise	0.826	0.985
Up	Disgust	0.735	0.906
Up	Fear	−0.880	0.974
Up	Happy	0.444	−0.588
Up	Surprise	−0.723	0.992

5. Discussion

The experimental results provide a bounded characterization of BGDC rather than evidence of uniform superiority across all metrics. On FER2013-VAD, BGDC achieves the highest mean classification accuracy and the highest mean V-A-D CCC, while Deep Ensemble remains stronger in ECE and MSE and has a slightly higher epistemic ranking. The FERPlus transfer experiment further shows that BGDC should not be interpreted as a general-purpose cross-dataset classifier since Deep Ensemble obtains stronger classification transfer performance. The main transferable property of BGDC is instead the preservation of an epistemic uncertainty signal that remains informative for detecting prediction errors. Taken together, these results characterize BGDC as a single-model reliability-aware FER framework that balances dimensional affective concordance and uncertainty-error alignment, rather than as a method that dominates every metric.

The ablation and sensitivity analyses further delimit the source of this behavior. The improvement in epistemic uncertainty-error alignment is mainly attributable to input-dependent posterior parameterization rather than to increased model size or to generic Bayesian sampling. The loss-weight sweeps also show that BGDC operates under competing objectives, so the final configuration should be interpreted as a task-level compromise rather than as an optimum for every individual metric.

The controlled masking experiment qualifies a common assumption about Bayesian-form estimators. Prediction error generally increases as facial evidence is removed, but epistemic uncertainty does not necessarily provide a valid degradation signal for all uncertainty estimators. The key criterion is not whether uncertainty increases with masking ratio in every spatial condition but whether it responds to the loss of task-relevant facial evidence. BGDC satisfies this criterion under center, downward, and random masking, whereas upward masking acts as a boundary condition in which posterior contraction can occur when the removed region changes the context representation without proportionally re-

moving expression-defining evidence. This result also explains why MC Dropout and Deep Ensemble should be empirically validated under structured facial degradation rather than assumed to provide reliable epistemic signals by construction.

The aleatoric and epistemic components also need not change in the same direction under degradation. Aleatoric uncertainty captures observation-level ambiguity, whereas epistemic uncertainty reflects instability in the learned mapping. Their relative contributions depend on the type and location of facial evidence being removed. In BGDC, part of the reliability-related variation is carried by the epistemic branch when degradation affects expression-defining regions. This supports the need to analyze aleatoric and epistemic uncertainty jointly rather than treating either component as a universal reliability measure.

Several limitations remain. FER2013-VAD provides paired categorical and continuous V-A-D annotations, but it is limited in scale, label granularity, and distributional diversity. FERPlus provides useful external validation for classification and uncertainty-based error detection, but it does not provide the same three-dimensional V-A-D supervision and therefore cannot serve as a full cross-dataset V-A-D regression benchmark. The masking protocol is deliberately simple to support controlled interpretation, and it does not cover mixed real-world degradations such as blur, illumination shift, low resolution, pose variation, or natural occlusion. Extending the evaluation to additional datasets, compatible V-A-D annotations, and more naturalistic degradation conditions remains necessary.

In summary, the present results support a bounded but practically relevant conclusion. BGDC does not uniformly outperform all baselines across all predictive, calibration, and computational criteria. Its main contribution lies in combining categorical emotion recognition, V-A-D concordance, and an empirically informative epistemic uncertainty signal within a single-model framework. The additional ablation, sensitivity, transfer, calibration, and masking analyses further show that reliability-aware FER should be evaluated through multiple complementary criteria rather than through accuracy or uncertainty magnitude alone. These findings define the current scope of BGDC and also indicate the next necessary step: validation on more diverse affective datasets and more naturalistic degradation conditions with compatible dimensional emotion annotations.

6. Conclusions

This study investigated multi-task facial emotion recognition from the perspective of uncertainty-aware predictive reliability under controlled facial information loss. The proposed BGDC framework integrates GMM-based soft semantic consistency, a context-conditioned Bayesian regression head, and an input-dependent aleatoric branch within a unified architecture. A controlled masking protocol was used as a diagnostic intervention to examine whether predictive uncertainty remains aligned with facial evidence degradation.

The experimental findings lead to three main conclusions. First, on FER2013-VAD, BGDC achieves the highest mean classification accuracy and the highest mean V-A-D CCC among the compared single-model configurations while providing a more informative epistemic uncertainty-error correspondence than MC Dropout. Second, the FERPlus transfer experiment shows that BGDC is not the strongest cross-dataset classifier, but its epistemic uncertainty remains useful for identifying prediction failures under distributional transfer. Third, the controlled masking analysis shows that BGDC epistemic uncertainty increases under center, downward, and random masking, where the removed regions more often overlap expression-relevant facial evidence, whereas upward masking is a boundary case that must be interpreted with respect to which facial regions carry the affective evidence.

These findings support a shift from deterministic affect prediction toward reliability-aware affect inference. In this view, uncertainty is not an auxiliary output but an operational signal for identifying when facial evidence is insufficient for stable V-A-D prediction.

The empirical scope remains constrained by the availability of datasets with compatible categorical and continuous affect annotations. Future validation on additional datasets containing compatible V-A-D annotations is necessary to establish broader generalization of the dimensional affective modeling component. Future work will extend the framework to additional datasets, richer affective label structures, and more diverse real-world degradation conditions, with the broader goal of developing probabilistic multi-task emotion recognition into a foundation for trustworthy and interpretable affective computing systems.

Author Contributions: Conceptualization, Q.X. and C.Q.; methodology, Q.X.; software, Q.X.; validation, Q.X.; formal analysis, Q.X.; investigation, Q.X.; resources, C.Q.; data curation, Q.X.; writing—original draft preparation, Q.X.; writing—review and editing, C.Q.; visualization, Q.X.; supervision, C.Q.; project administration, C.Q.; funding acquisition, C.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Number JP25K15078.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed in this study is publicly available from the GitHub repository at https://github.com/YeeHoran/Multi-task-Emotion-Recognition (accessed on 21 June 2026) . The processed experimental results supporting the findings of this study are available from the corresponding author upon reasonable request, subject to the license and usage restrictions of the original dataset.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

FER	Facial Emotion Recognition
V-A-D	Valence–Arousal–Dominance
BGDC	Bayesian Gaussian-mixture Distributional Consistency
GMM	Gaussian Mixture Model
MC	Monte Carlo
OOD	Out-of-Distribution
ECE	Expected Calibration Error
CCC	Concordance Correlation Coefficient
MSE	Mean Squared Error
NLL	Negative Log-Likelihood
VI	Variational Inference
KL	Kullback-Leibler divergence

Appendix A. Additional Per-Sample Masking Visualizations

This appendix provides additional qualitative visualizations for the controlled masking analysis in Section 4.6. These examples are not used as independent statistical tests. Their purpose is to illustrate how the same masking ratio can correspond to different degrees of affect-relevant evidence loss depending on the facial expression category and the spatial direction of the mask.

Together with the Happy example shown in Figure 7, Figures A1 and A2 show that the semantic effect of masking is not determined by the masking ratio alone. In the Happy sample, downward masking progressively removes the mouth and lower-face

region, whereas upward masking first affects the hat and upper forehead. For the Sad sample, center masking covers the central facial configuration, while downward masking progressively removes the open-mouth region. For the Surprise sample, the eye region is visually prominent, so upward and center masking can affect diagnostically relevant upper-face evidence more directly than in the Happy example. These examples provide visual support for the region-dependent interpretation used in the main text.

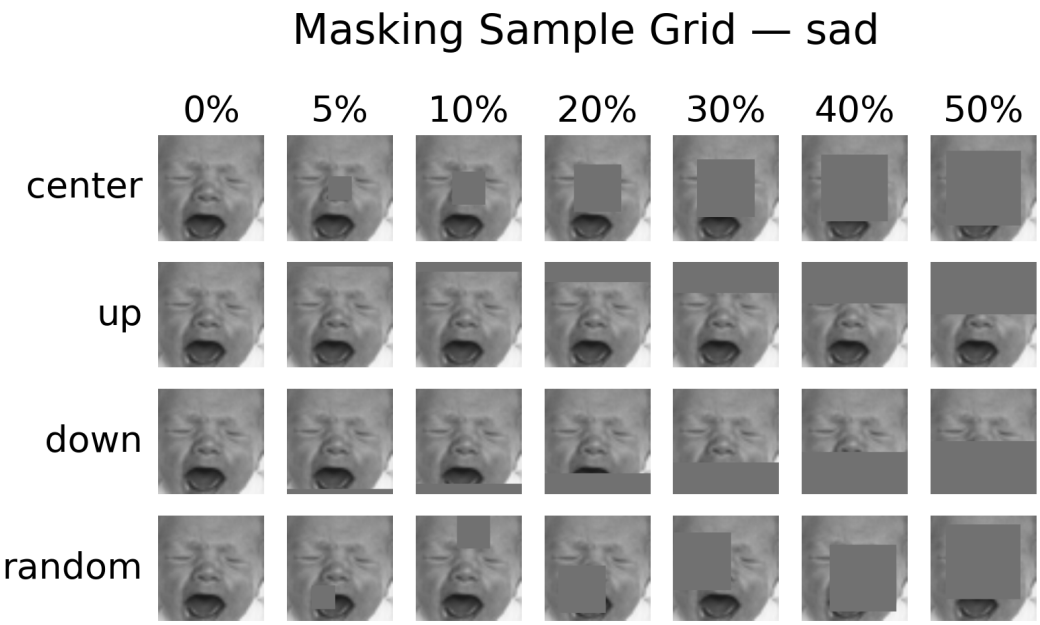


Figure A1. Masking grid for the representative Sad sample. Columns correspond to masking ratios from 0% to 50%, and rows correspond to center, upward, downward, and random masking.

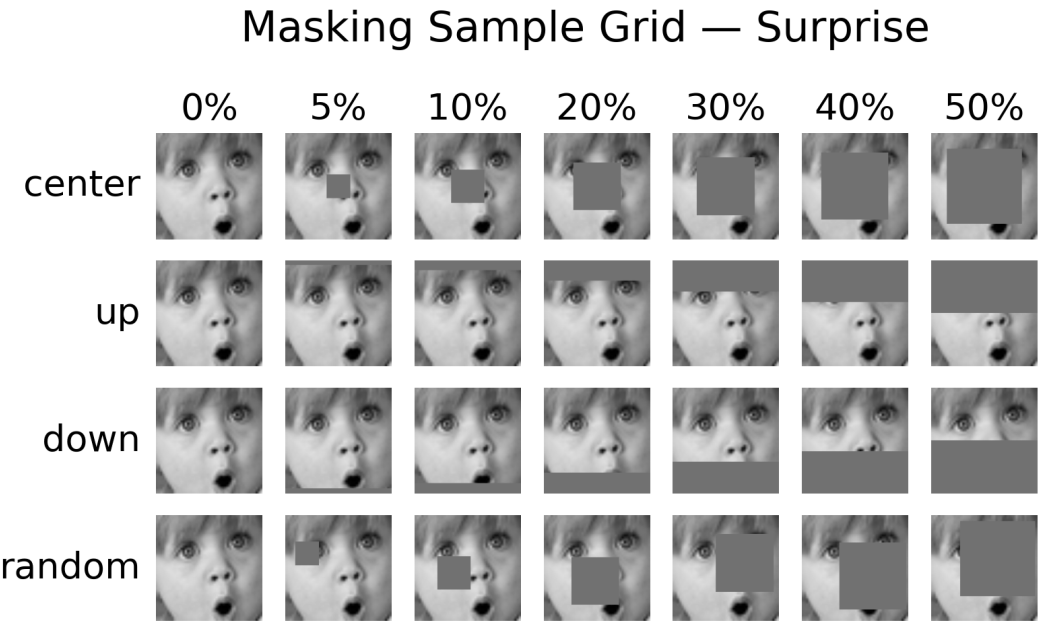


Figure A2. Masking grid for the representative Surprise sample. Columns correspond to masking ratios from 0% to 50%, and rows correspond to center, upward, downward, and random masking.

Appendix B. Matrix Visualization of Per-Sample Uncertainty Responses

Figure A3 provides a compact view of how epistemic uncertainty changes with prediction error at the individual-sample level. The figure shows that the uncertainty response is

not a uniform function of masking ratio. Instead, it depends on whether the masked region overlaps facial evidence that is relevant to the corresponding expression.

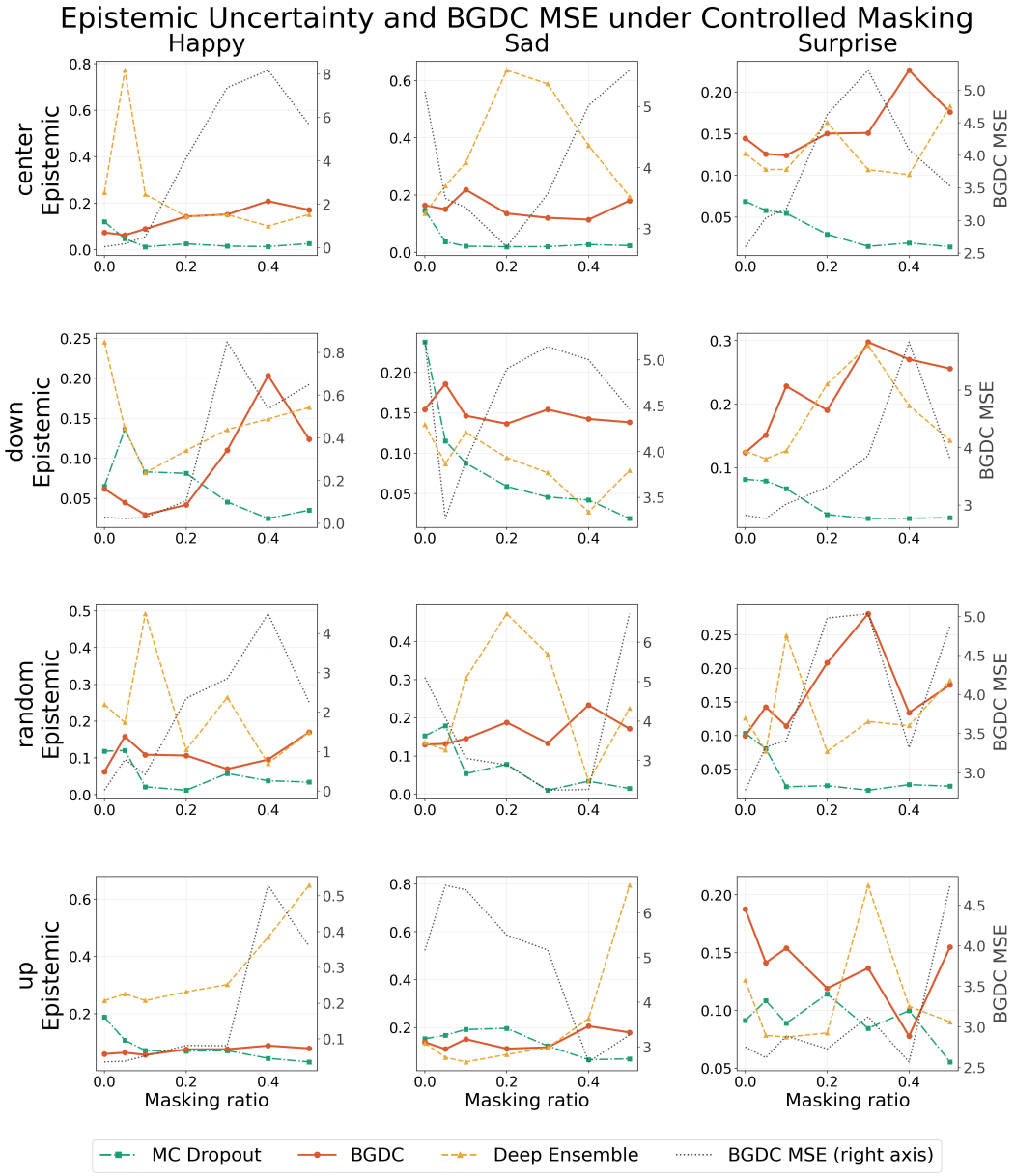


Figure A3. Epistemic uncertainty and BGDC MSE under controlled masking for representative Happy, Sad, and Surprise samples. Rows correspond to masking directions, columns correspond to emotion samples, and the dotted black curve denotes BGDC MSE on the right axis.

For the Happy sample, the matrix confirms the interpretation given in the main text. Under center masking, BGDC MSE increases sharply after the mask begins to cover the central facial structure, and BGDC epistemic uncertainty increases from approximately 0.07 at the unmasked condition to above 0.20 at the 40% ratio. Under downward masking, the lower-face evidence is progressively removed, and BGDC epistemic uncertainty also rises in the middle-to-late masking stages. In contrast, upward masking produces a much weaker BGDC epistemic response because the early and intermediate upward masks mainly remove the hat and upper forehead rather than the smile-related mouth region. This supports the claim that upward masking is a boundary case rather than standalone evidence of improved reliability.

For the Sad sample, the response is less monotonic at the individual-sample level, but it still shows that masking direction changes the relation between uncertainty and prediction error. Under center masking, Deep Ensemble produces a large local epistemic peak at intermediate ratios, whereas BGDC remains more moderate. Under downward masking, MC Dropout decreases from a high initial epistemic value to a much lower value as the masking ratio increases, even though the lower facial region is progressively degraded. This behavior is consistent with the aggregate result that MC Dropout can produce decreasing epistemic variance under structured evidence removal. BGDC does not produce a strictly monotonic per-sample curve for this Sad instance, but its response remains more stable than the sharp contraction observed for MC Dropout.

For the Surprise sample, center and downward masking provide clearer evidence of region-dependent uncertainty behavior. Under center masking, BGDC epistemic uncertainty increases at higher masking ratios when the mask interferes with the central facial configuration, and the BGDC MSE also rises in the intermediate stages. Under downward masking, BGDC epistemic uncertainty increases strongly from the unmasked condition to the 30–40% ratios, where the mask begins to remove expression-relevant lower-face structure. Random masking produces more irregular curves because the mask location changes across ratios and may overlap either informative regions or less relevant background-like areas. Therefore, the random case should not be interpreted as a simple degradation schedule, but as a spatially variable perturbation.

Across the three samples, MC Dropout frequently shows decreasing epistemic uncertainty as the masking ratio increases, even when prediction error remains high or increases. Deep Ensemble often produces large local peaks, but these peaks are not consistently aligned with masking severity or BGDC MSE. BGDC generally gives a more region-dependent response: epistemic uncertainty tends to increase when center, downward, or random masking removes expression-relevant evidence, while upward masking produces weaker or contracting responses when the removed region is less directly tied to the dominant affective cue. This pattern is consistent with the aggregate monotonicity analysis in Table 8 and the per-emotion summary in Table 9.

References

1. Picard, R.W. *Affective Computing*; The MIT Press: Cambridge, MA, USA, 1997.
2. Kalateh, S.; Estrada-Jimenez, L.A.; Nikghadam Hojjati, S.; Barata, J. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access* **2024**, *12*, 103976–104019. [CrossRef]
3. Huo, Y.; Ge, Y. Multi-Task Emotion Recognition Based on Dimensional Model and Category Label. *IEEE Access* **2024**, *12*, 75169–75179. [CrossRef]
4. Xiao, Q.; Quan, C.; Cao, S.; Luo, Z. Gaussian Mixture Model-Based Soft Consistency for Multi-Task Emotion Recognition. *IEEE Access* **2025**, *13*, 183690–183705. [CrossRef]
5. Kollias, D.; Sharmanska, V.; Zafeiriou, S. Distribution matching for multi-task learning of classification tasks: A large-scale study on faces & beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2024; Volume 38, pp. 2813–2821.
6. Park, S.; Kim, J.; Ye, S.; Jeon, J.; Park, H.Y.; Oh, A. Dimensional emotion detection from categorical emotion. *arXiv* **2019**, arXiv:1911.02499.
7. Wagner, N.; Mätzler, F.; Vossberg, S.R.; Schneider, H.; Pavlitska, S.; Zöllner, J.M. Cage: Circumplex affect guided expression inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 4683–4692.
8. Wan, W.; Yu, C.; Chen, J.; Wu, T.; Zhong, Y.; Yang, M.H. Shaping deep feature space towards gaussian mixture for visual classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2430–2444. [CrossRef]
9. Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; Ding, G. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Trans. Multimed.* **2016**, *19*, 632–645. [CrossRef]
10. Neto, J.B.C.; Ferrari, C.; Berretti, S. Revisiting Emotions Representation for Recognition in the Wild. *arXiv* **2026**, arXiv:2602.06778.

11. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017.
12. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16), New York, NY, USA, 20–22 June 2016*; Journal of Machine Learning Research: Norfolk, MA, USA, 2016; Volume 48, pp. 1050–1059.
13. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight uncertainty in neural network. In *Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015*; Journal of Machine Learning Research: Norfolk, MA, USA, 2015; pp. 1613–1622.
14. Kingma, D.P.; Salimans, T.; Welling, M. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2015.
15. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017.
16. Liu, J.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax Weiss, T.; Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems 33 (NIPS 2020)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2020; pp. 7498–7512.
17. Wang, K.; Cuzzolin, F.; Shariatmadar, K.; Moens, D.; Hallez, H. A Review of Uncertainty Representation and Quantification in Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *48*, 2476–2495. [CrossRef]
18. She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; Mei, T. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 6248–6257.
19. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 6897–6906.
20. Zhang, Y.; Wang, C.; Deng, W. Relative uncertainty learning for facial expression recognition. In *Advances in Neural Information Processing Systems 34 (NIPS 2021)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2021; pp. 17616–17627.
21. Le, N.; Nguyen, K.; Tran, Q.; Tjiputra, E.; Le, B.; Nguyen, A. Uncertainty-aware label distribution learning for facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023*; IEEE: New York, NY, USA, 2023; pp. 6088–6097.
22. Liu, Y.; Zhang, X.; Kauttonen, J.; Zhao, G. Uncertain facial expression recognition via multi-task assisted correction. *IEEE Trans. Multimed.* **2023**, *26*, 2531–2543. [CrossRef]
23. Matin, M.; Valdenegro-Toro, M. Hey human, if your facial emotions are uncertain, you should use bayesian neural networks! *arXiv* **2020**, arXiv:2008.07426.
24. Cohen, I.; Sebe, N.; Gozman, F.; Cirelo, M.C.; Huang, T.S. Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June 2003*; IEEE: New York, NY, USA, 2003; Volume 1.
25. Wang, Y.; Yan, S.; Liu, Y.; Song, W.; Liu, J.; Chang, Y.; Mai, X.; Hu, X.; Zhang, W.; Gan, Z. A survey on facial expression recognition of static and dynamic emotions. *arXiv* **2024**, arXiv:2408.15777.
26. Dapogny, A.; Bailly, K.; Dubuisson, S. Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *Int. J. Comput. Vis.* **2018**, *126*, 255–271.
27. Li, Y.; Liu, H.; Liang, J.; Jiang, D. Occlusion-Robust Facial Expression Recognition Based on Multi-Angle Feature Extraction. *Appl. Sci.* **2025**, *15*, 5139. [CrossRef]
28. Cortiñas-Lorenzo, K.; Lacey, G. Toward explainable affective computing: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 13101–13121. [CrossRef]
29. Ong, D.C.; Soh, H.; Zaki, J.; Goodman, N.D. Applying probabilistic programming to affective computing. *IEEE Trans. Affect. Comput.* **2019**, *12*, 306–317. [CrossRef] [PubMed]
30. Rohanian, M.; Hüppi, R.; Nooralahzadeh, F.; Dannecker, N.; Pauli, Y.; Surbeck, W.; Sommer, I.; Hinzen, W.; Langer, N.; Krauthammer, M.; et al. Uncertainty modeling in multimodal speech analysis across the psychosis spectrum. *npj Digit. Med.* **2026**, *9*, 218. [CrossRef] [PubMed]
31. Chen, H.; Zendehdel, N.; Leu, M.C.; Yin, Z. Fine-grained activity classification in assembly based on multi-visual modalities. *J. Intell. Manuf.* **2023**, *35*, 2215–2233. [CrossRef]
32. Li, Y.; Liu, H.; Jiang, D.; Liang, J. Facial Expression Recognition with Label-Noisy Under Dual-Branch Noise Extraction and Suppression. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1514–1525.

33. Lo, L.; Ruan, B.K.; Shuai, H.H.; Cheng, W.H. Modeling Uncertainty for Low-Resolution Facial Expression Recognition. *IEEE Trans. Affect. Comput.* **2024**, *15*, 198–209. [CrossRef]
34. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 279–283.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.