Neural Mixture Density Processes

Yi Ding, Qi Tao, Xingxing Liang*, Longfei Zhang, Yiqin Lv, Weitao Song, Fangjie Yang,
Qi Wang*, Guangquan Cheng*
National University of Defense Technology
doublestar_l@163.com, hhq123go@gmail.com, cgq299@nudt.edu.cn

Abstract

The neural process (NP) is a probabilistic meta-learning model that learns distributions over functions via a global latent variable. It enables fast adaptation in few-shot scenarios by leveraging past experience. However, the design of latent variable structures and conditioning mechanisms in NPs remains underexplored, despite their importance in capturing diverse functional distributions. This paper proposes a new variant of NPs via mixture density modeling, referred to as the neural mixture density process (NMDP). The NMDP decomposes model parameters into task-agnostic and task-specific components to represent function distributions more flexibly. We train the model using a variational EM/MM-style procedure with self-normalized importance sampling, yielding an explicit surrogate objective for learning expressive functional priors. Compared with existing work, our method maintains several advantages: (i) efficient adaptation at test time by only inferring a compact task-specific latent variable, (ii) compact task representation via distributions in the simplex, (iii) a principled EM/MM-style optimization with a monotonic-improvement guarantee in the idealized exact-inference setting. Experimental results show that our method can achieve competitive performance with adequate explainability.

1. Introduction

In recent years, probabilistic meta-learning has emerged as a compelling paradigm in the machine learning community [10, 12, 13, 53, 56]. Leveraging past experience, it learns to adapt to new tasks rapidly while providing uncertainty quantification in prediction. In this work, we focus on a representative line of probabilistic meta-learning models, stochastic process–based approaches, with particular attention to the Neural Process (NP) framework [14]. NPs approximate the functional prior of the underlying task distribution by representing the context dataset into a global latent variable, serving as a scalable alternative to Gaussian Processes (GPs) [51] with reduced computational cost.

However, existing variants of NPs [13, 14, 25] typically assume Gaussian likelihoods, thereby restricting the model to unimodal predictive distributions. Such an assumption limits their expressiveness when modeling inherently multimodal function behaviors that frequently arise in real-world tasks. In these cases, classical NP design fails to capture the uncertainty structure and diversity of possible outputs conditioned on the same context set.

In order to capture arbitrarily complex function distributions, we present a novel variant of NPs, termed the Neural Mixture Density Process (NMDP). The NMDP builds on the classical mixture density network framework [2], representing functional uncertainty through a latent variable z defined on an L-dimensional simplex Δ^{L-1}. The architecture decouples into a task-agnostic component, a set of probabilistic density estimators, and a task-specific component that infers the functional prior from a few-shot data points. For tractable optimization, we derive an importance-weighted variational objective inspired by the reweighted wake-sleep method [3], securing stable convergence and enriching expressiveness of the underlying function space.

Outline and Contributions. Section 2 overviews related work, followed by the background of NPs in Section 3. Section 4 presents the proposed NMDP, its optimization framework, and the meta-training pipeline. The subsequent sections report experimental evaluations, analyses, and concluding remarks. Primarily, our contributions are threefold:

1. *Novel neural process formulation.* We introduce the NMDP, an exchangeable stochastic process model designed to scale over mixture function distributions. Drawing inspiration from Dirichlet processes [50], the NMDP can theoretically approximate arbitrarily flexible distributions over functions.
2. *Principled optimization objective.* Departing from conventional variational inference–based training of latent variable models [14, 27, 36], we derive an importance-weighted objective that enables robust learning of functional priors within mixture distributions.

*Corresponding Authors.

3. *Comprehensive empirical validation.* We construct benchmarks involving diverse and complex function families, demonstrating that the NMDP achieves (i) competitive predictive performance, (ii) well-calibrated uncertainty estimation, and (iii) interpretable function representations.

2. Literature Review

Meta-Learning. The essence of meta-learning lies in devising mechanisms that extract transferable knowledge from prior experience to facilitate rapid adaptation to new tasks [1, 15, 16, 18, 19, 21, 22, 32, 42, 43, 57, 58, 61, 62]. Such an approach alleviates the need to relearn similar tasks from scratch, thereby improving data and computational efficiency. Existing meta-learning frameworks can broadly be categorized into optimization-based, context-based, and metric-based methods. Optimization-based approaches, exemplified by Model-Agnostic Meta-Learning (MAML) [6, 9, 34, 44, 63, 64], formulate the problem as a bi-level optimization task solved via gradient-based updates. Context-based methods, such as the NP family [13, 14, 21, 25, 40, 59], instead learn task representations that enable fast adaptation within a probabilistic framework. Metric-based models, typified by the Prototypical Network [30, 47], focus on learning embedding spaces that preserve task-relevant similarity structures, particularly for few-shot image classification.

Neural Process Family. NPs provide a probabilistic meta-learning framework in which a global latent variable encodes the context set as functional prior, facilitating rapid task adaptation [7, 24, 33, 35, 46, 55, 60]. However, standard NPs often exhibit underfitting, motivating a range of methodological extensions to enhance representational and inferential capacity. Attention-based variants [25, 26] improve predictive accuracy through introducing a local deterministic embedding for each data point. Convolutional formulations such as ConvCNP [17] and ConvNP [11] incorporate translation equivariance into NP modules for generalization improvement, and mixture-of-experts approaches [56] employ context-conditioned expert selection to model heterogeneous task distributions. Geometry-aware regularization has also been explored to refine uncertainty calibration through bi-Lipschitz constraints [52]. Distinguished from prior works that mainly introduce structural inductive biases into Neural Processes, our approach focuses on designing NP modules capable of approximating more complex stochastic processes and developing tractable inference strategies for their effective optimization.

3. Preliminaries

Notations. Let $p(\tau)$ denote a distribution over functions, where each sampled function is represented by τ . We define the context set $\mathcal{D}_\tau^C = \{(x_i, y_i)\}_{i=1}^n$ for learning functional priors and the target set $\mathcal{D}_\tau^T = \{(x_i, y_i)\}_{i=1}^{n+m}$ for prediction. The latent variable z inferred from $\mathcal{D}_\tau^C$ serves as a compact representation of the underlying function. This work considers the distribution over functions as an exchangeable stochastic process ($\mathcal{SP}$), defined in Appendix A.

Neural Processes. The standard NPs introduce a global latent variable that constitutes a family of $\mathcal{SP}$. The corresponding generative model is expressed as:

$$\rho_{x_{1:n}}(y_{1:n}) = \int p(z) \prod_{i=1}^{n} \mathcal{N}\left(y_i; \mu_\theta([x_i, z]), \Sigma_\theta([x_i, z])\right) dz, \tag{1}$$

where the input to the generative network is the concatenation of the latent variable z and the index variable x_i. The functions $\{\mu_\theta, \Sigma_\theta\}$ parameterize the mean and covariance of the predictive Gaussian distribution, respectively.

Approximate Inference. As the exact forms of the functional prior and posterior are generally intractable, variational inference is employed to optimize an approximate evidence lower bound (ELBO) for the NPs, as defined in Eq. (2):

$$\begin{aligned} \mathcal{L}_{\text{NP}}(\theta, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(z|\mathcal{D}_\tau^T)} \left[\ln p(\mathcal{D}_\tau^T|z;\theta)\right] \\ -D_{KL}\left[q_{\boldsymbol{\phi}}(z|\mathcal{D}_\tau^T) \parallel q_{\boldsymbol{\phi}}(z|\mathcal{D}_\tau^C)\right]. \end{aligned} \tag{2}$$

Here, the approximate posterior $q_{\boldsymbol{\phi}}(z|\mathcal{D}_\tau^T)$ and the approximate prior $q_{\boldsymbol{\phi}}(z|\mathcal{D}_\tau^C)$ share the same neural structure. The target observations are assumed to be conditionally independent given the latent variable such that $p(\mathcal{D}_\tau^T|z;\theta) = \prod_{i=1}^{n+m} p(y_i|x_i, z;\theta)$.

4. Methodology

This section details a novel exchangeable $\mathcal{SP}$, referred to as the proposed NMDP, formulates the approximate optimization objective with inference strategies, and presents some theoretical analysis.

4.1. Neural Mixture Density Process

The type of global latent variables and the way to condition them in generative processes are key design choices in the NP family. Motivated by this insight, we propose the NMDP framework, in which a set of probability density functions is shared across all tasks, and the latent variable z, constrained to a simplex Δ^{L-1}, governs the mixture over expert components.

In other words, the NMDP defines a family of explicit mixture distributions for meta-learning. Here, η and ψ denote the parameters of the prior inference network $p(z|\mathcal{D}_\tau^C;\eta)$ and the generative network $p(\mathcal{D}_\tau^T|z;\psi)$, respectively. In particular, ψ comprises the parameters of a set of probabilistic density networks, represented as

$\{\psi_1, \psi_2, \ldots, \psi_L\}$. The purpose of meta-training is to discover the shared probabilistic components $\{p_{\psi_l}(\cdot|\cdot)\}_{l=1}^L$ that best explain the meta-learning datasets and uncover the clustering property among tasks.

Generative Processes. As depicted in Fig. 1, the probabilistic neural module comprises a Dirichlet latent variable $z \in \Delta^{L-1}$ and a set of L probabilistic density networks with deterministic parameters $\psi = \{\psi_l\}_{l=1}^L$. Accordingly, the generative process for a single function τ can be characterized as follows.

$$\rho_{x_{1:n}}(y_{1:n} \mid \psi) = \int \mathsf{Dir}(z;\alpha) \prod_{i=1}^{n} \left[\sum_{l=1}^{L} z_l \, p_{\psi_l}(y_i|x_i) \right] dz. \tag{3}$$

Eq. (3) shares a form similar to Eq. (1) in NPs, but the generative distribution over functions in the NMDP framework exhibits greater expressiveness, as highlighted in **Remark** 1. Importantly, the latent variable z in NMDPs linearly modulates the different probabilistic components. An illustrative example is provided below.

Example 1 (Gaussian Neural Mixture Density Processes) *Consider the task-relevant latent variable $z \sim \mathsf{Dir}_\eta(z;\alpha)$ and a collection of Gaussian density experts $p_{\psi_l}(y|x) = \mathcal{N}(y; \mu_{\psi_l}(x), \Sigma_{\psi_l}(x))$. Then the conditional predictive density of a Gaussian NMDP takes the form*

$$p(y|x,z) = \sum_{l=1}^{L} z_l \, \mathcal{N}(y; \mu_{\psi_l}(x), \Sigma_{\psi_l}(x)).$$

Remark 1 *In principle, the proposed NMDPs can approximate arbitrary function distributions when sufficiently many probabilistic components $\{p_{\psi_l}(\cdot|\cdot)\}_{l=1}^L$ are used to span the task distribution [28, 39].*

Optimization Objectives. We aim to maximize the likelihood of the conditional distribution $p(\mathcal{D}_{\mathcal{T}}^T|\mathcal{D}_{\mathcal{T}}^C)$ in meta-learning, where $\mathcal{T}$ is a batch of sampled tasks. Moreover, by introducing latent variables, this likelihood can be factorized as shown in Eq. (4).

$$p(\mathcal{D}_{\mathcal{T}}^T|\mathcal{D}_{\mathcal{T}}^C) = \prod_{\tau \in \mathcal{T}} \left[\int p(\mathcal{D}_\tau^T|z;\psi) p(z|\mathcal{D}_\tau^C;\eta) dz \right]. \tag{4}$$

Here $p(z|\mathcal{D}_\tau^C;\eta) = \mathsf{Dir}_\eta(z;\alpha)$ defines the Dirichlet functional prior with the concentration parameter α, which is implicitly conditioned on the context data points $\mathcal{D}_\tau^C$. The generative likelihood of target data points can be rewritten as

$$p(\mathcal{D}_\tau^T|z;\psi) = \prod_{i=1}^{n+m} \left[\sum_{l=1}^{L} z_l p_{\psi_l}(y_i|x_i) \right] \tag{5}$$

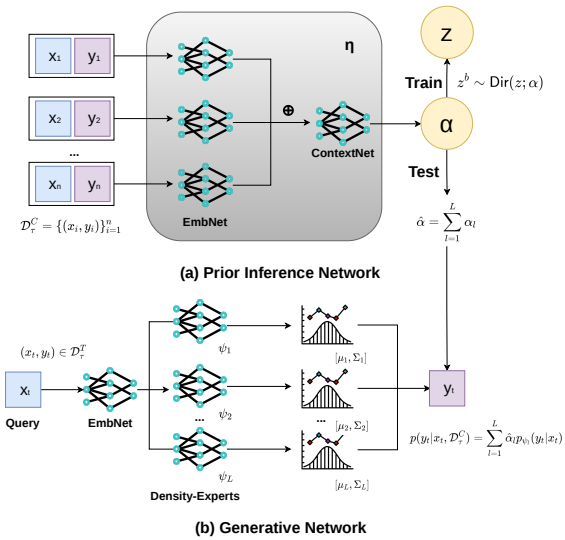


Figure 1. **Neural Architectures of the NMDP.** The prior inference network is a permutation invariant and task-specific network to parameterize a Dirichlet distribution $\mathsf{Dir}(z;\alpha)$. The generative module corresponds to a collection of probability density networks with z to formulate mixture distributions.

In Bayesian inference, the posterior plays a crucial role. With the Dirichlet functional prior $p(z|\mathcal{D}_\tau^C;\eta) = \mathrm{Dir}_\eta(z;\alpha)$ and the generative distribution $p(\mathcal{D}_\tau^T|z;\psi)$ in the NMDP, the posterior can be written in the form

$$p(z|\mathcal{D}_\tau^T, \mathcal{D}_\tau^C) = \frac{\mathrm{Dir}_\eta(z;\alpha) \prod_{i=1}^{n+m} \left[\sum_{l=1}^{L} z_l p_{\psi_l}(y_i|x_i) \right]}{\int \mathrm{Dir}_\eta(z;\alpha) \prod_{i=1}^{n+m} \left[\sum_{l=1}^{L} z_l p_{\psi_l}(y_i|x_i) \right] dz}. \tag{6}$$

Note that the derived functional posterior is generally intractable because the denominator involves a non-analytical integral.

In the following sections, we consider a single task $\tau \in \mathcal{T}$ to formulate equations for ease of presentation, but the model is trained over a batch of tasks during meta-learning.

4.2. Tractable Optimization with Importance Sampling

In the same way as Eq. (2), one may derive an ELBO for the log-likelihood $\mathcal{L}(\eta, \psi) := \ln p(\mathcal{D}_\tau^T|\mathcal{D}_\tau^C)$ associated with Eq. (4). However, such an approach relies on an auxiliary variational distribution and may introduce a posterior approximation gap.

Instead, we optimize an EM/MM-style surrogate objective inspired by the reweighted wake-sleep algorithm [3]. Let $\Theta = (\eta, \psi)$ denote the model parameters and $\Theta_t = (\eta_t, \psi_t)$ denote the parameters from the previous outer iteration. The resulting surrogate objective for the conditional marginal likelihood is formulated as follows:

$$\max_{\eta,\psi} \mathcal{L}(\eta,\psi) = \mathbb{E}_{p(z|\mathcal{D}_\tau^T,\mathcal{D}_\tau^C;\Theta_t)}\left[\ln p(\mathcal{D}_\tau^T, z|\mathcal{D}_\tau^C;\Theta)\right]$$
$$= \mathbb{E}_{p(z|\mathcal{D}_\tau^T,\mathcal{D}_\tau^C;\Theta_t)}\left[\ln \underbrace{p(\mathcal{D}_\tau^T|z;\psi)}_{\text{Mixture Distribution}} + \ln \underbrace{p(z|\mathcal{D}_\tau^C;\eta)}_{\text{Dirichlet Prior}}\right]. \tag{7}$$

It can be observed that Eq. (7) comprises a generative mixture distribution term and a Dirichlet functional prior term, and the distribution inside the expectation is treated as fixed when optimizing terms inside the bracket.

Remark 2 *Eq. (7) defines the ideal EM surrogate objective of the NMDP. Under exact posterior evaluation and exact surrogate maximization, the corresponding EM/MM procedure recovers the standard monotonic-improvement property.*

Since it is intractable to directly sample from the posterior $p(z|\mathcal{D}_\tau^T, \mathcal{D}_\tau^C; \Theta_t)$, we cannot derive a Monte Carlo estimate of Eq. (7). To address this, we employ a self-normalized importance sampling strategy, using the current prior from the previous outer iteration $p(z|\mathcal{D}_\tau^C;\eta_t)$ as the proposal distribution to generate inference particles. In this way, we can formulate an importance-weighted optimization objective $\mathcal{L}_{\mathsf{IW}}(\eta,\psi)$ as follows.

$$\max_{\eta,\psi} \mathcal{L}_{\mathsf{IW}}(\eta,\psi) = \mathbb{E}_{p(z|\mathcal{D}_\tau^C;\eta_t)}\left[\omega_t(z)\ln p(\mathcal{D}_\tau^T, z|\mathcal{D}_\tau^C;\Theta)\right]$$
$$\approx \mathcal{L}_{\mathsf{IW\text{-}MC}}(\eta,\psi;B) = \sum_{b=1}^{B} \hat{\omega}_t^{(b)} \ln p(\mathcal{D}_\tau^T, z^{(b)}|\mathcal{D}_\tau^C;\Theta)$$
$$= \sum_{b=1}^{B} \hat{\omega}_t^{(b)} \Big[\ln p(z^{(b)}|\mathcal{D}_\tau^C;\eta) + \ln p(\mathcal{D}_\tau^T|z^{(b)};\psi)\Big], \tag{8}$$

where the (unnormalized) importance weights satisfy $\omega_t(z) \propto p(\mathcal{D}_\tau^T \mid z;\psi_t)$ under the proposal $p(z|\mathcal{D}_\tau^C;\eta_t)$. The inference particles are drawn B times from the proposal distribution $z^{(b)} \sim p(z|\mathcal{D}_\tau^C;\eta_t)$, and the self-normalized importance weights are computed via:

$$\omega_t^{(b)} = p(\mathcal{D}_\tau^T \mid z^{(b)};\psi_t),$$
$$\hat{\omega}_t^{(b)} = \frac{\omega_t^{(b)}}{\sum_{b'=1}^{B}\omega_t^{(b')}} = \frac{p(\mathcal{D}_\tau^T \mid z^{(b)};\psi_t)}{\sum_{b'=1}^{B} p(\mathcal{D}_\tau^T \mid z^{(b')};\psi_t)}. \tag{9}$$

Please refer to Appendix C for more derivation details about the above equations.

4.3. Training and Prediction

Given the above optimization objective, the NMDP is instantiated with neural modules comprising a context-conditioned Dirichlet prior network and a mixture-of-experts generative network, as depicted in Fig. 1; complete architectural details are provided in Appendix D.2.

Algorithm 1: Meta-Training NMDPs.

Input : Task distribution $p(\tau)$; Batch size $|\mathcal{T}|$; Number of particles B; Learning rate λ; Total iterations N_{iter}.
Output : Meta-trained model parameters η and ψ.

Initialize the model parameters η_0 and ψ_0;
for $t = 0$ ***to*** $N_{iter} - 1$ **do**
  // **E-step: Estimate importance weights**
  Initialize gradients: $\nabla_\eta\mathcal{L} \leftarrow 0, \nabla_\psi\mathcal{L} \leftarrow 0$;
  for *each task* $\tau \in \mathcal{T}$ *(a batch of tasks)* **do**
    Perform sampling to obtain particles $z^{(b)} \sim p(z|\mathcal{D}_\tau^C;\eta_t)$ for $b = 1,\ldots,B$;
    Compute the self-normalized importance weights $\{\hat{\omega}^{(b)}\}_{b=1}^{B}$ via Eq. (9);
    Accumulate gradients for this task: $\nabla_\eta\mathcal{L}$ += $\sum_{b=1}^{B}\hat{\omega}^{(b)}\nabla_\eta \ln p(z^{(b)}|\mathcal{D}_\tau^C;\eta)$;
    $\nabla_\psi\mathcal{L}$ += $\sum_{b=1}^{B}\hat{\omega}^{(b)}\nabla_\psi \ln p(\mathcal{D}_\tau^T|z^{(b)};\psi)$;
  end
  // **M-step: Update meta model parameters**
  Average gradients over the batch: $\nabla_\eta\mathcal{L} \leftarrow \nabla_\eta\mathcal{L}/|\mathcal{T}|$, $\nabla_\psi\mathcal{L} \leftarrow \nabla_\psi\mathcal{L}/|\mathcal{T}|$;
  Execute gradient updates: $\eta_{t+1} \leftarrow \eta_t + \lambda\nabla_\eta\mathcal{L}$;
  $\psi_{t+1} \leftarrow \psi_t + \lambda\nabla_\psi\mathcal{L}$;
end

Iterative Steps in Meta-Training. As shown in Algorithm 1, the meta-training process iteratively optimizes the model parameters over a batch of tasks $\mathcal{T}$. In each iteration, it consists of two main steps: (i) **E-step**, where for each task we sample particles from the current functional prior $p(z|\mathcal{D}_\tau^C;\eta_t)$ and compute importance weights; and (ii) **M-step**, where we update the parameters η and ψ by maximizing the importance-weighted surrogate objective averaged across the task batch. Fig. 2 illustrates the optimization process in detail.

By iterating these steps until convergence, the procedure follows an approximate EM/MM-style surrogate optimization rationale, where the posterior expectation is approximated by self-normalized importance sampling and the surrogate is optimized by gradient-based updates, as discussed in Remark 2.

Predictive Distribution. During meta-testing, the context set $\mathcal{D}_\tau^C$ is fed into the prior inference network to induce the Dirichlet functional prior $p(z|\mathcal{D}_\tau^C;\eta) = \mathsf{Dir}_\eta(z;\alpha)$. With the meta-trained NMDP's parameters η and ψ, the predictive distribution of a new data point (x_*, y_*) can be written with a closed-form expression by marginalizing out

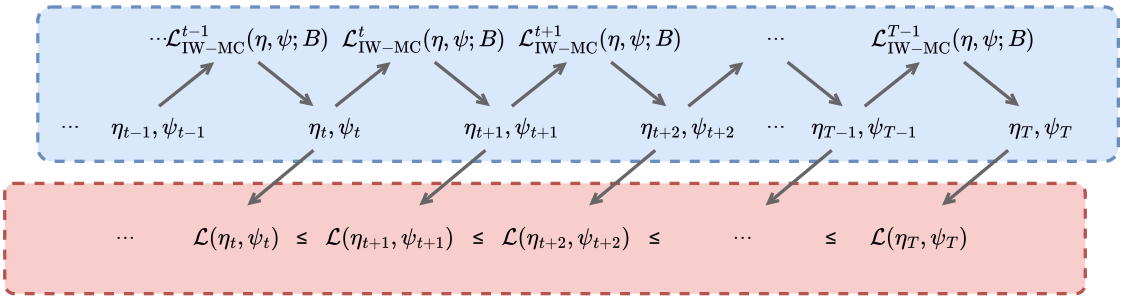


Figure 2. **Optimization Process with Surrogate Functions for Neural Mixture Density Processes.** The red block depicts the iterative refinement of the surrogate objective. The blue block illustrates the EM/MM-style surrogate optimization rationale *w.r.t.* the marginal likelihood $\mathcal{L}(\eta_t, \psi_t)$ under the learned context-conditioned Dirichlet prior. From left to right, the surrogate is constructed using the previous parameters $\{\eta_t, \psi_t\}$ and then approximately optimized by maximizing $\mathcal{L}_{\text{IW-MC}}(\eta, \psi; B)$. T denotes the maximum number of inner optimization steps used to approximately optimize the surrogate at each outer iteration.

the latent variables as follows:

$$p(y_*|x_*, \mathcal{D}_\tau^C; \eta, \psi) = \int \text{Dir}_\eta(z;\alpha) \left[\sum_{l=1}^{L} z_l p_{\psi_l}(y_*|x_*)\right] dz$$
$$= \sum_{l=1}^{L} \frac{\alpha_l}{\sum_{l'=1}^{L} \alpha_{l'}} p_{\psi_l}(y_*|x_*) = \sum_{l=1}^{L} \hat{\alpha}_l p_{\psi_l}(y_*|x_*), \tag{10}$$

where $[\alpha_1, \ldots, \alpha_L]$ and $[\hat{\alpha}_1, \ldots, \hat{\alpha}_L]$ denote the concentration and mean parameters of the learned Dirichlet functional prior, respectively.

4.4. Understanding NMDPs

Model Design. Here we interpret the proposed NMDP as a finite-mixture analogue of Dirichlet-process-style mixture modeling [50]. When the number of probabilistic components is conceptually extended to the infinite limit, the model connects to the perspective of random probability measures over expert parameters. Specifically, let $\mathcal{G}_\mathcal{T} = \{G_\tau : \tau \in \mathcal{T}\}$ denote a collection of probability measures on the parameter space Ψ, induced by the distribution of meta-learning tasks. Given a data point $(x, y) \in \mathcal{D}_\tau^T$ and the context points $\mathcal{D}_\tau^C$ for a task τ, the NMDP can be viewed as an infinite weighted combination of component distributions in this limiting sense. Taking the Gaussian NMDP as an instance, a family of Gaussian mixture models is induced as follows:

$$p(y_{1:n+m} \mid x_{1:n+m}, \mathcal{D}_\tau^C)$$
$$= \int_\Psi \prod_{i=1}^{n+m} p(y_i \mid x_i, \psi_\ell)\, G_\tau(d\psi_\ell), \tag{11}$$
$$\text{with} \quad p_{\psi_\ell}(y_i \mid x_i) = \mathcal{N}\big(y_i; \mu_{\psi_\ell}(x_i), \Sigma_{\psi_\ell}(x_i)\big),$$

where, in the infinite-component limit, we can write $G_\tau = \sum_{\ell=1}^{\infty} z_{\tau,\ell}\, \delta_{\psi_{\tau,\ell}}$ with $\delta_{\psi_{\tau,\ell}}$ denoting a Dirac measure and $\{z_{\tau,\ell}\}_{\ell=1}^{\infty}$ denoting mixture weights. The corresponding simplified expression is $p(y|x, \mathcal{D}_\tau^C) = \sum_{\ell=1}^{\infty} z_{\tau,\ell}\, \mathcal{N}(y; \mu_{\tau,\ell}(x), \Sigma_{\tau,\ell}(x))$. Leveraging the meta-knowledge that the base distribution $\mathcal{G}_\mathcal{T}$ is shared across tasks in our setting, we impose the constraint $\delta_{\psi_{\tau,\ell}} := \delta_{\psi_\ell}$ in modeling. This interpretation is mainly conceptual and serves to motivate the flexibility of the proposed finite-mixture NMDP.

Optimization Strategies. The derived optimization objective is inspired by the reweighted wake-sleep (RWS) algorithm [3]. The primary difference from the original RWS algorithm is that we learn a permutation-invariant, context-conditioned Dirichlet prior, whereas the prior is typically fixed in RWS. Furthermore, by reusing the current prior as the proposal distribution, our optimization avoids introducing an additional sleep-phase inference network. Overall, the resulting procedure is best understood as an approximate EM/MM-style optimization algorithm for NMDP.

5. Experiments and Analysis

Unlike prior work, which mainly pursues state-of-the-art (SOTA) performance from a structural inductive bias perspective, we take more interest in answering the following research questions (RQs):

1. *Is the Dirichlet functional prior more expressive in modeling complicated function distributions?*
2. *Can the NMDP reveal the clustering properties inside the task distribution and provide better task representations?*
3. *Does the variational expectation maximization algorithm work better than variational inference in optimizing the NMDP?*

We evaluated the proposed method on two representative scenarios: one-dimensional (1D) synthetic function regression, two-dimensional (2D) image completion and multi-input-output regression. Comparison baselines include context-based methods CNP [13], ANP [25], ConvCNP [17], TNP [40] and DNP [52], as well as gradient-based methods MAML [9] and CAVIA [64]. All models employ identical Gaussian likelihood, data normalization, and implementation settings to ensure comparability.

5.1. 1D Function Regression on Synthetic GP Data

To evaluate NMDP's ability to model heterogeneous function distributions, we design synthetic regression tasks by sam-

Table 1. Average log-likelihood on synthetic 1D regression tasks (mean ± standard error over 1,000 tasks).

Model	RBF	Weakly Periodic	Matérn$-\frac{5}{2}$
MAML [9]	0.08±1e-3	0.89±3e-3	-0.12±1e-3
CAVIA [64]	0.21±2e-3	0.96±1e-3	0.07±3e-3
CNP [13]	0.15±4e-3	1.08±2e-3	-0.14±2e-3
ANP [25]	0.42±2e-3	1.11±4e-3	0.18±2e-3
ConvCNP [17]	1.05±5e-3	0.80±2e-3	0.71±4e-3
TNP [40]	1.23±4e-3	1.13±3e-3	1.06±5e-3
DNP [52]	0.95±6e-3	1.06±3e-3	0.78±3e-3
NMDP (Ours)	**1.26**±4e-3	**1.19**±2e-3	**1.15**±1e-3

pling functions from Gaussian processes with diverse characteristics. In contrast to conventional approaches that train separate models for each kernel-specific data distribution [17], we train on a mixture of functions drawn from multiple kernels. Specifically, the cross-kernel mixture dataset is constructed by sampling each task uniformly from one of three distinct GP priors: the exponentiated quadratic (EQ/RBF) kernel, a Weakly Periodic kernel, and the Matérn$-\frac{5}{2}$ kernel. This design explicitly introduces cross-modal statistical heterogeneity, requiring models to simultaneously capture smooth, rough, and periodic function behaviors. At each training iteration, context $\mathcal{D}_\tau^C = \{(x_i, y_i)\}_{i=1}^n$ and target $\mathcal{D}_\tau^T = \{(x_i, y_i)\}_{i=1}^{n+m}$ points are randomly selected from the sampled functions within the interval $[-2, 2]$. During meta-testing, we evaluate performance over 1,000 independently generated tasks using the same sampling procedure.

Table 1 summarizes quantitative performance across all competing methods. NMDP achieves state-of-the-art results on all three kernel types, demonstrating consistent advantages over both gradient-based meta-learners (MAML, CAVIA) and neural process variants (CNP, ANP, ConvCNP, TNP, DNP). Moreover, NMDP maintains more stable performance across different kernel types, indicating superior generalization to heterogeneous function families. Fig. 3 provides qualitative comparisons of predictive distributions. Compared to CNP and ANP, NMDP produces predictions that more closely align with ground truth functions while providing well-calibrated uncertainty estimates. These results directly address our first research question (RQ-1): by introducing a Dirichlet functional prior and task-level mixture density modeling, NMDP demonstrates enhanced expressiveness and robust cross-kernel generalization on complex, multi-modal function families.

Task Representation and Visualization. To assess whether NMDP captures meaningful task-level representations aligned with underlying functional structures, we extract Dirichlet concentration vectors from the inferred priors and apply centered log-ratio (CLR) transformation [23]. These embeddings are visualized using UMAP [37], and their structure is quantitatively evaluated in the original CLR space using normalized mutual information (NMI) [49], adjusted rand index (ARI) [20], silhouette coefficient [45], and linear probe accuracy [4] to prevent over-interpretation of low-dimensional projections.

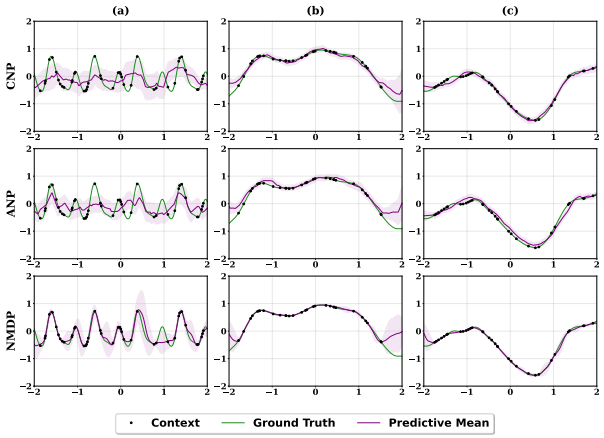


Figure 3. **Predictive performance comparison of three neural process variants: CNP (top row), ANP (middle row), and NMDP (bottom row).** Models were trained on synthetic data generated from a heterogeneous mixture of Gaussian process priors, including RBF, Matérn$-\frac{5}{2}$, and weakly periodic kernels. The green curve depicts the underlying ground truth function, while the black points represent the empirical observations. The purple curve illustrates the model's posterior predictive mean, and the shaded regions delineate the ±1 standard deviation confidence interval.

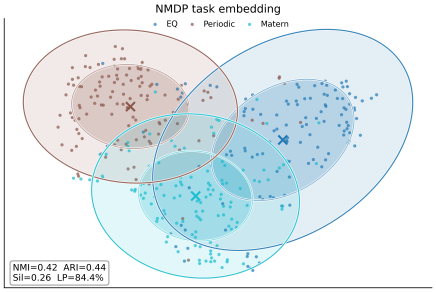
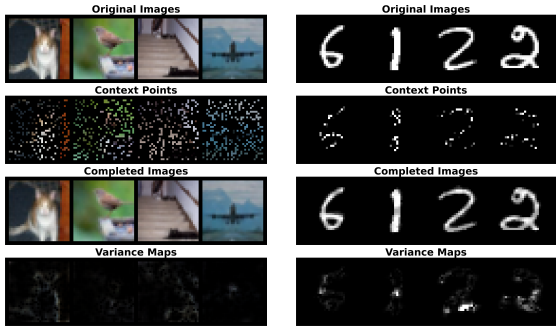


Figure 4. **Task embeddings generated by the NMDP model when learning different function types.** Each point represents a task, with colors distinguishing three distinct function types: RBF, Weakly Periodic, and Matern. Dirichlet concentration vectors from the inferred prior are transformed with a centered log-ratio (CLR) and projected to 2D with UMAP.

As shown in Fig. 4, NMDP yields distinct clusters corresponding to RBF, Weakly Periodic, and Matérn functions, with NMI = 0.42, ARI = 0.44, Silhouette = 0.26, and linear-probe accuracy of 84.4% (chance ≈ 33%). The strong linear separability confirms that kernel-type information is reliably encoded, while moderate unsupervised scores suggest partial overlap among Gaussian-based kernels due to shared statistical properties.

The above findings address RQ-2, confirming that the NMDP model effectively identifies cross-task correlations and produces interpretable task representations that reflect the true functional diversity of the underlying processes.



(a)CIFAR-10 (b)MNIST

Figure 5. **Image completion results across different datasets.** Each subplot from top to bottom shows the original image, context points, reconstructed output and predictive variance.

5.2. Image Completion

To assess the efficacy of NMDP on high-dimensional visual tasks with inherent uncertainty, we conduct image completion experiments on four widely-adopted benchmark datasets: CIFAR-10 [29], SVHN [38] (3-channel RGB images), and MNIST [31], EMNIST [5] (1-channel grayscale images). Following established protocols [14, 25], we formulate each image as a continuous function $f : [-1, 1]^2 \rightarrow [0, 1]^c$ that maps 2D pixel coordinates $\mathbf{x}$ to normalized intensity values $\mathbf{y}$ (where $c = 3$ for RGB or $c = 1$ for grayscale). During meta-training, given a context set comprising observed pixel locations and their corresponding values, the model outputs predictive Gaussian distribution parameters (mean and variance) for all pixel coordinates, thereby enabling probabilistic image reconstruction under uncertainty. Fig. 5 presents representative completion results across datasets.

Table 2 summarizes quantitative performance across all competing methods, measured by average test log-likelihood over 10 independent runs. On 3-channel RGB datasets (CIFAR-10 and SVHN), NMDP substantially outperforms all baselines on both context and target predictions. Specifically, on CIFAR-10, our method achieves target log-likelihood of 4.19, representing a 3.9% improvement over the strongest baseline (TNP: 4.03). This superior performance on structurally complex, multi-channel natural images validates NMDP's enhanced capacity to capture intricate color dependencies and high-dimensional correlations. On single-channel grayscale datasets (MNIST and EMNIST), NMDP remains highly competitive, achieving the best target log-likelihood (1.42 for MNIST, 1.17 for EMNIST) despite TNP obtaining marginally higher context log-likelihood. The smaller performance gap on these structurally simpler tasks (grayscale digits with less inter-channel dependency) suggests that the additional modeling capacity of NMDP becomes particularly advantageous when handling complex multi-channel correlations inherent in natural RGB images.

These results collectively validate that our probabilistic mixture modeling approach is particularly effective for complex, high-dimensional visual reconstruction tasks where output distributions exhibit substantial multi-modality. Comprehensive implementation details are provided in the Appendix D.

5.3. Multi-input-output Regression

To assess the efficacy of the NMDP in more challenging and realistic scenarios, we follow the evaluation protocols established in [52, 55] and benchmark its performance on three widely used multivariate regression datasets: SARCOS [54], Water Quality (WQ) [8], and SCM20D [48]. The SARCOS dataset comprises 48,933 samples, each with 21 input features and 7 target outputs. The WQ dataset consists of 1,060 instances with 16 input variables and 14 outputs, whereas SCM20D contains 8,966 samples characterized by 61 inputs and 16 outputs. Model performance is quantified using Mean Squared Error (MSE) for predictive accuracy.

As summarized in Table 3, NMDP demonstrates consistently strong predictive performance and outperforms baseline methods in the majority of cases.

5.4. Ablation Studies

To dissect the contribution of each component in our proposed NMDP, we conduct a five-stage ablation study on the synthetic Gaussian process (GP) benchmark under the protocol outlined in Section 5.1. Each variant incrementally incorporates a new modeling element, allowing us to isolate its specific contribution:

- **A1 (Baseline):** A single-expert model without mixture components ($L = 1$), serving as our simplest baseline.
- **A2 (Uniform Mixture):** A multi-expert model with fixed, uniform mixture weights ($L > 1$, fixed z), introducing decoder diversity while remaining non-adaptive.
- **A3 (Deterministic Gating):** Extends A2 by replacing uniform weights with learnable, deterministic weights generated by a context encoder, allowing adaptive expert selection.
- **A4 (Stochastic Gating, ELBO):** Further introduces a Dirichlet latent variable z governing the mixture weights, optimized through the standard ELBO objective to capture task-specific uncertainty.
- **A5 (Full Model, NMDP):** The complete NMDP framework refines inference by optimizing the Dirichlet latent variable with the tighter importance-weighted objective.

Progressive performance analysis. As shown in Fig. 6, validation log-likelihood improves monotonically across variants A1–A5, confirming the incremental value of each architectural refinement. The baseline single-expert model (A1) establishes a lower performance bound; introducing decoder diversity via a uniform mixture (A2) yields modest gains, while adaptive gating through a context encoder (A3) further

Table 2. Average Log-likelihood from image experiments (10 runs). We test the performance of different models in both context data points and target data points.

Model	SVHN		CIFAR10		MNIST		EMNIST	
	context	target	context	target	context	target	context	target
MAML [9]	2.15±6e-3	2.03±5e-3	1.97±3e-3	1.81±6e-3	0.88±5e-3	0.79±3e-3	0.77±8e-3	0.71±1e-3
CAVIA [64]	2.88±6e-3	2.69±5e-3	3.01±4e-3	2.85±7e-3	1.06±2e-3	0.97±1e-3	0.94±3e-3	0.88±2e-3
CNP [13]	2.77±6e-3	2.62±6e-3	2.63±4e-3	2.51±3e-3	1.09±2e-3	1.01±2e-3	1.02±6e-3	0.86±5e-3
ANP [25]	3.22±4e-3	3.05±4e-3	3.95±6e-3	3.76±3e-3	1.14±3e-3	1.03±3e-3	1.05±3e-3	0.92±2e-3
ConvCNP [17]	3.13±8e-3	3.08±7e-3	4.03±9e-3	3.88±6e-3	1.26±3e-3	1.17±4e-3	1.24±4e-3	1.12±3e-3
TNP [40]	3.28±5e-3	3.23±7e-3	4.17±6e-3	4.03±8e-3	**1.51**±7e-3	1.38±4e-3	**1.29**±5e-3	1.13±3e-3
DNP [52]	3.24±3e-3	3.10±4e-3	4.02±7e-3	3.81±6e-3	1.20±2e-3	1.09±3e-3	1.11±2e-3	0.95±2e-3
NMDP (Ours)	**3.41**±7e-3	**3.29**±5e-3	**4.33**±9e-3	**4.19**±6e-3	1.47±4e-3	**1.42**±5e-3	1.25±3e-3	**1.17**±3e-3

Table 3. Average Testing MSEs on Multi-Output Dataset (10 runs)

Model	SARCOS	WQ	SCM20D
MAML [9]	0.88±8e-3	0.77±4e-3	0.92±7e-3
CAVIA [64]	0.84±6e-3	0.73±5e-3	0.91±4e-3
CNP [13]	0.95±7e-3	0.74±9e-3	0.95±5e-3
ANP [25]	1.01±3e-2	0.69±5e-3	0.93±6e-3
ConvCNP [17]	0.94±8e-3	0.78±9e-2	0.82±8e-3
TNP [40]	0.91±6e-3	0.70±3e-3	0.84±9e-3
DNP [52]	0.88±7e-3	0.71±5e-3	0.89±4e-3
NMDP (Ours)	**0.82**±3e-3	**0.67**±6e-3	**0.75**±5e-3

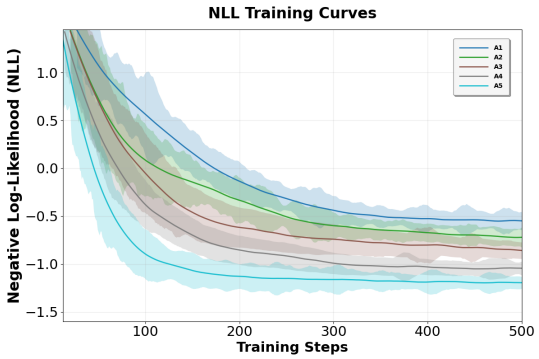


Figure 6. **Ablation Study Results.** Validation negative log-likelihood on the synthetic GP dataset for the five model variants (A1-A5). Shaded regions represent ±3 standard error over multiple independent runs.

enhances expressivity by enabling input-dependent expert selection.

The most substantial improvements arise from our core innovations in A4 and A5. In A4, we replace deterministic gating with a Dirichlet-distributed latent variable z that explicitly models uncertainty over task identity, optimized via the standard evidence lower bound (ELBO). This stochastic formulation proves especially advantageous under sparse-context regimes, where ambiguity in function type necessitates probabilistic reasoning over expert assignments. Finally, the full NMDP model A5 adopts an importance-weighted objective to refine posterior inference over z. This yields not only higher asymptotic performance but also markedly reduced training variance, as evidenced by narrower confidence intervals across runs. Crucially, the tighter variational bound circumvents the approximation gap inherent in ELBO-based methods, leading to more faithful optimization of the functional prior and enhanced stability in meta-learning. Moreover, convergence accelerates progressively from A1 to A5, with the full model achieving stable optima significantly faster, indicative of improved gradient signal quality and parameter efficiency.

Collectively, these progressive improvements establish that mixture capacity, adaptive gating, and our proposed stochastic task representation with importance-weighted optimization act synergistically to confer superior performance on multi-modal function families.

6. Discussion and Conclusion

Summary. We proposed Neural Mixture Density Processes (NMDP), a neural process variant that represents task-level uncertainty by mixing a shared set of density experts with Dirichlet simplex weights inferred from context. We train NMDP with an importance-weighted surrogate objective, approximated efficiently via Monte Carlo self-normalized importance sampling. Experiments on synthetic regression, image completion, and multi-output regression show that NMDP yields competitive predictive accuracy, improved uncertainty calibration on heterogeneous multi-modal tasks, and interpretable task representations through learned mixture weights.

Limitations. When the test task is far from the meta-training distribution or the context is insufficiently informative, the inferred Dirichlet prior can become high-entropy (near-uniform), leading to averaged expert usage and increased predictive uncertainty; in practice, Dirichlet entropy may serve as a simple ambiguity indicator. NMDP is most effective when cross-task variation is well captured by a finite set of shared experts. Importance-weighted training incurs additional computation due to multiple particles, though this cost is largely parallelizable on modern GPUs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) with the Number # 62306326, 72301289 and 62433021.

References

[1] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International conference on artificial intelligence and statistics*, pages 670–678. PMLR, 2018. 2

[2] Christopher M Bishop. Mixture density networks. 1994. 1

[3] Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014. 1, 3, 5

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 6

[5] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017. 7, 4

[6] Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 33:18860–18871, 2020. 2

[7] Hongkun Dou, Junzhe Lu, Zeyu Li, Xiaoqing Zhong, Wen Yao, Lijun Yang, and Yue Deng. Score-based neural processes. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 2

[8] Sašo Džeroski, Damjan Demšar, and Jasna Grbović. Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, 13(1):7–17, 2000. 7

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 2, 5, 6, 8

[10] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018. 1

[11] Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems*, 33:8284–8295, 2020. 2

[12] Meijun Fu, Xiaomin Wang, Jun Wang, and Zhang Yi. Generative probabilistic meta-learning for few-shot image classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 1

[13] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018. 1, 2, 5, 6, 8

[14] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018. 1, 2, 7

[15] Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir H Gandomi. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Computing Surveys*, 56(12):1–41, 2024. 2

[16] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*. 2

[17] Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2019. 2, 5, 6, 8, 4

[18] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations*, 2018. 2

[19] Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[20] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. 6

[21] Ekaterina Iakovleva, Jakob Verbeek, and Karteek Alahari. Meta-learning with shared amortized variational inference. In *International Conference on Machine Learning*, pages 4572–4582. PMLR, 2020. 2

[22] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11719–11727, 2019. 2

[23] M. Chris Jones and John William Aitchison. The statistical analysis of compositional data. 1986. 6

[24] Makoto Kawano, Wataru Kumagai, Akiyoshi Sannai, Yusuke Iwasawa, and Yutaka Matsuo. Group equivariant conditional neural processes. *arXiv preprint arXiv:2102.08759*, 2021. 2

[25] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019. 1, 2, 5, 6, 7, 8

[26] Mingyu Kim, Kyeongryeol Go, and Se-Young Yun. Neural processes with stochastic attention: Paying more attention to the context dataset. *arXiv preprint arXiv:2204.05449*, 2022. 2

[27] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 1

[28] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. *arXiv preprint arXiv:2206.10044*, 2022. 3

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7, 4

[30] Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, 34:24581–24592, 2021. 2

[31] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002. 7, 4

[32] Xiaoxu Li, Zhuo Sun, Jing-Hao Xue, and Zhanyu Ma. A concise review of recent few-shot meta-learning methods. *Neurocomputing*, 456:463–468, 2021. 2

[33] Huafeng Liu, Yiran Fu, Liping Jing, Hui Li, Shuyang Lin, Jingyue Shi, Deqiang Ouyang, and Jian Yu. Learning robust neural processes with risk-averse stochastic optimization. In *Forty-second International Conference on Machine Learning*. 2

[34] Yiqin Lv, Qi Wang, Dong Liang, and Zheng Xie. Theoretical investigations and practical enhancements on tail task risk minimization in meta learning. *Advances in Neural Information Processing Systems*, 37:82921–82961, 2024. 2

[35] Yiqin Lv, Dong Liang, Wumei Du, Zenglin Shi, Zheng Xie, Qi Wang, and Meng Wang. Tail task risk minimization in meta-learning from theoretical advances to practical strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 2

[36] Jacqueline Maasch, Willie Neiswanger, Stefano Ermon, and Volodymyr Kuleshov. Probabilistic graphical models: A concise tutorial. *arXiv preprint arXiv:2507.17116*, 2025. 1

[37] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 6

[38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 7. Granada, 2011. 7, 4

[39] Hien D Nguyen and Geoffrey McLachlan. On approximations via convolution-defined mixture models. *Communications in Statistics-Theory and Methods*, 48(16):3945–3955, 2019. 3

[40] Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In *International Conference on Machine Learning*, pages 16569–16594. PMLR, 2022. 2, 5, 6, 8

[41] Bernt Øksendal. Stochastic differential equations, stochastic differential equations, 2003. 1

[42] Massimiliano Patacchiola, Jack Turner, Elliot J Crowley, Michael O'Boyle, and Amos J Storkey. Bayesian meta-learning for the few-shot setting via deep kernels. *Advances in Neural Information Processing Systems*, 33:16108–16118, 2020. 2

[43] Yun Qu, Qi Cheems Wang, Yixiu Mao, Yiqin Lv, and Xiangyang Ji. Fast and robust: Task sampling with posterior and diversity synergies for adaptive decision-makers in randomized environments. *arXiv preprint arXiv:2504.19139*, 2025. 2

[44] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019. 2

[45] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 6

[46] Jiayi Shen, Xiantong Zhen, Qi Wang, and Marcel Worring. Episodic multi-task learning with heterogeneous neural processes. *Advances in Neural Information Processing Systems*, 36:75214–75228, 2023. 2

[47] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 2

[48] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98, 2016. 7

[49] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002. 6

[50] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006. 1, 5

[51] Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 844–851. JMLR Workshop and Conference Proceedings, 2010. 1

[52] Aishwarya Venkataramanan and Joachim Denzler. Distance-informed neural processes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 5, 6, 7, 8

[53] Anna Vettoruzzo, Mohamed-Rafik Bouguelia, Joaquin Vanschoren, Thorsteinn Rögnvaldsson, and KC Santosh. Advances and challenges in meta-learning: A technical review. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):4763–4779, 2024. 1

[54] Sethu Vijayakumar and Stefan Schaal. Locally weighted projection regression: An o (n) algorithm for incremental real time learning in high dimensional space. In *Proceedings of the seventeenth international conference on machine learning (ICML 2000)*, pages 288–293. Morgan Kaufmann Burlington, MA, 2000. 7

[55] Qi Wang and Herke Van Hoof. Doubly stochastic variational inference for neural processes with hierarchical latent variables. In *International Conference on Machine Learning*, pages 10018–10028. PMLR, 2020. 2, 7

[56] Qi Wang and Herke van Hoof. Learning expressive meta-representations with mixture of expert neural processes. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[57] Qi Wang and Herke Van Hoof. Model-based meta reinforcement learning using graph structured surrogate models and amortized policy search. In *International Conference on Machine Learning*, pages 23055–23077. PMLR, 2022. 2

[58] Qi Wang, Yiqin Lv, Zheng Xie, Jincai Huang, et al. A simple yet effective strategy to robustify the meta learning paradigm. *Advances in Neural Information Processing Systems*, 36:12897–12928, 2023. 2

[59] Qi Wang, Marco Federici, and Herke van Hoof. Bridge the inference gaps of neural processes via expectation maximization. *arXiv preprint arXiv:2501.03264*, 2025. 2

[60] Qi Wang, Yiqin Lv, Yixiu Mao, Yun Qu, Yi Xu, and Xiangyang Ji. Robust fast adaptation from adversarially explicit task distribution generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1481–1491, 2025. 2

[61] Qi Wang, Zehao Xiao, Yixiu Mao, Yun Qu, Jiayi Shen, Yiqin Lv, and Xiangyang Ji. Model predictive task sampling for efficient and robust adaptation. *arXiv preprint arXiv:2501.11039*, 2025. 2

[62] Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, and Yee Whye Teh. Metafun: Meta-learning with iterative functional updates. In *International Conference on Machine Learning*, pages 10617–10627. PMLR, 2020. 2

[63] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7343–7353, 2018. 2

[64] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019. 2, 5, 6, 8