Dual-Process Distribution Calibration: Bridging Slow-Fast Thinking for Few-Shot Learning

Yuchen Liu[1,2], **Weining Weng**[1,2], **Lingxing Chen**[1,2], **Qianzhong Chen**[1,2], **Shiyang Li**[1,2], **Yuan Ma**[1,2], **Yang Gu**[1,2*]

[1]Beijing Key Laboratory for Multimodal Collaboration and Advanced Application, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
liuyuchen23s@ict.ac.cn, wengweining21b@ict.ac.cn, chenlingxing19@mails.ucas.ac.cn, chenqianzhong24s@ict.ac.cn, lishiyang24s@ict.ac.cn, mayuan20z@ict.ac.cn, guyang@ict.ac.cn

Abstract

Artificial intelligence models typically perform well on large-scale datasets, yet their effectiveness tends to degrade in real-world scenarios with scarce data, such as medical diagnostics. In contrast, humans can learn and reason effectively from few examples. Even when novel objects differ significantly from prior observations, humans can quickly infer category distributions accurately. Inspired by this, we explore how to leverage human-inspired cognitive mechanisms to improve the distribution calibration in few-shot learning models. Based on the "fast-slow thinking" dual-process theory, we propose a novel cognition-inspired few-shot learning framework. It mimics the human cognition when handling novel information: it first rapidly screens relevant knowledge through "fast thinking" and then infers inter-class relationships and calibrates distributions via "slow thinking". Specifically, the fast-thinking stage employs a gating mechanism to quickly match input samples with known base-class prototypes, activating relevant candidate knowledge. In the slow-thinking stage, the model aggregates inter-class edge embeddings into a summary relation graph and then applies divergent Gaussian sampling to generate multiple relation graphs representing different association strengths, thereby achieving distribution calibration for few-shot classes. Experiments on public few-shot benchmarks and medical image datasets show competitive performance. Visualization analyses further reveal that our framework exhibits meaningful interpretability.

1 Introduction

In real-world scenarios such as medical diagnosis and rare species identification, acquiring abundant labeled data is often difficult or costly. Few-shot learning enables models to recognize new categories from extremely limited samples. Existing few-shot learning approaches fall into three paradigms: fine-tuning, metric learning, and data augmentation [Wang *et al.*, 2020]. Among them, data augmentation methods are commonly adopted to mitigate the fundamental difficulty of accurately estimating the underlying category distribution from sparse samples [Yang *et al.*, 2021], as they can effectively enrich the feature distributions of few-shot classes by synthesizing additional training instances [Sun *et al.*, 2025].

To better estimate and calibrate feature-space distributions, researchers have proposed various data augmentation methods. These methods expand diversity at the feature-space level to address sample scarcity, for example by leveraging unlabeled data [Ren *et al.*, 2018] or applying inter-class statistical distances for distribution calibration [Yang *et al.*, 2021]. Recent work also explores causal intervention strategies that remove potential bias paths to optimize distributional rationality [Wu *et al.*, 2025]. However, a fundamental challenge remains: augmented data distributions must avoid excessive concentration or drift from the true distribution to ensure effective calibration in complex few-shot tasks [Wang *et al.*, 2020]. Notably, humans demonstrate high cognitive efficiency, rapidly extracting core features from few samples through associative reasoning and naturally grasping category distributions for robust judgments [Pennycook *et al.*, 2015]. This insight inspires us: if data augmentation models could also "think like humans", mimicking their synergistic process of rapid screening and deep reasoning, we might enhance the effectiveness of distribution calibration and open new avenues for few-shot learning.

To investigate how few-shot learners can approximate human-like reasoning under scarce supervision, we draw on dual-process theory from cognitive science. Prior studies suggest that humans rely on two complementary cognitive systems when making decisions from limited evidence, namely fast thinking and slow thinking [Kahneman, 2011]. Fast System 1 performs intuitive pattern matching and retrieves task-relevant prior knowledge to form preliminary hypotheses about inter-class associations. Slow System 2 further refines these hypotheses through deliberate integration and reasoning, explores multiple relational perspectives, and supports distribution-level inference from few observations.

Motivated by this cognitive mechanism, we introduce dual-process theory into few-shot learning and propose the **D**ual-**P**rocess **D**istribution **C**alibration **(DPDC)** framework. DPDC calibrates few-shot category distributions through two col-

laborative stages. In the fast-thinking stage, a gated prototype matching mechanism rapidly activates relevant base-class knowledge and produces an initial knowledge allocation over candidate slow-thinking components. In the slow-thinking stage, the model aggregates inter-class relations to construct a summary relation graph that constrains the distribution space, and then performs divergent Gaussian sampling to generate multiple relation graphs with different association strengths for distribution calibration. The main contributions of DPDC are summarized as follows:

- We propose a **dual-process distribution calibration framework** that models few-shot distribution estimation through fast knowledge matching and slow relational reasoning, thereby improving the accuracy and robustness of distribution calibration under data-scarce conditions.
- We develop an **interpretable relation graph modeling and divergent sampling mechanism** that reflects cognitive integration and divergence in human associative inference. The relation graphs explicitly encode inter-class associations, and divergent sampling derives diverse yet constrained relational hypotheses for calibrated feature generation.
- Experiments on multiple few-shot benchmarks and medical image datasets demonstrate that DPDC achieves superior accuracy over representative baselines. Visualization analyses further show that the **inter-class relation graphs** derived during slow thinking capture semantically meaningful associations and provide an intuitive rationale for model decisions.

By introducing theories from cognitive science, this work provides a principled framework for few-shot distribution calibration and offers insights into building more interpretable and human-inspired few-shot learning systems.

2 Related Works

Few-shot learning tackles the problem of identifying new classes with only a handful of labeled samples per category. Early work established several cornerstone approaches. Matching Networks [Vinyals *et al.*, 2016] employ attention over embedded support instances to classify queries. MAML [Finn *et al.*, 2017] learns parameter initializations that enable rapid adaptation to novel tasks. Prototypical Networks [Snell *et al.*, 2017] represent each class by its mean support embedding and perform nearest-neighbor classification. Distribution calibration [Yang *et al.*, 2021] adjusts novel-class statistics using base-class statistics to correct distributional bias.

Recent advances address challenges such as feature misalignment, background noise, and limited sample diversity in fine-grained or complex domains. [Alam *et al.*, 2025] improved feature extraction and alignment through cross-layer refinement and cross-sample adjustment. [Liu *et al.*, 2024] employed multi-scale features with spatial attention to select discriminative regions and suppress background interference via cross-interaction. [Li *et al.*, 2025] increased sample diversity through random geometric transformations and measured cross-view similarity between original and augmented features. [Zhou *et al.*, 2025b] learned optimal hyperparameters for transductive inference directly from validation data using an unrolled Expectation-Maximization strategy.

Concurrently, another research direction draws inspiration from cognitive science to build more robust and interpretable few-shot learning systems. These approaches mimic cognitive mechanisms to improve distribution estimation from sparse data. For example, [Wang *et al.*, 2024] structured visual knowledge using a Gaussian Mixture Model to emulate human intuitive clustering and memory. [Zhou *et al.*, 2025a] fused visual features with semantic language embeddings, reflecting the co-evolution of visual and linguistic understanding in human development. These works collectively highlight that the human cognitive system provides a powerful theoretical blueprint for mitigating distribution bias. Building upon the human dual-process cognitive system, our work proposes a framework that computationally instantiates this theory for few-shot distribution calibration.

3 Method

Inspired by the dual-process cognitive theory of human "fast and slow thinking", this paper presents the **Dual-Process Distribution Calibration (DPDC)** framework. DPDC simulates the cognitive pathway through which humans infer and calibrate category distributions from scarce information. As illustrated in Figure 1, the DPDC framework consists of three core parts: (1) The **fast-thinking component** acts as an efficient filter, rapidly screening and activating the base-class knowledge modules most relevant to the current task through gated prototype matching; (2) The **slow-thinking component** serves as the core reasoning engine, performing deep relation aggregation and multi-perspective divergent sampling on the activated knowledge to generate calibrated multi-perspective feature representations; (3) In the **feature integration and model optimization** stage, multi-perspective features from different slow-thinking components are fused, a lightweight classifier performs classification, and a joint loss function optimizes the entire system.

3.1 Fast-Thinking Component: Gated Prototype Matching and Dynamic Knowledge Allocation

Corresponding to the fast, intuitive System 1 in human cognition, our fast-thinking component simulates a rapid screening process. It identifies the most relevant knowledge from a large prior base for the current few-shot task. It centers on **Gated Prototype Matching**. Input features are extracted from few-shot episodes by a pre-trained backbone. The component contains a Gating Network and T independent slow-thinking components, each storing a set of base-class prototypes.

For an n-way k-shot task, let the support set features be $\mathcal{H}_S = \{\mathbf{h}_i^s \in \mathbb{R}^d\}_{i=1}^{n\times k}$. The gating network computes relevance between input features and knowledge across all slow-thinking components, generating an initial matching score g_t for each component t:

$$g_t = \bar{\mathbf{h}}_S^\top \mathbf{W}_g \mathbf{v}_t, \tag{1}$$

where $\bar{\mathbf{h}}_S$ denotes the mean vector of the support set features, $\mathbf{W}_g$ is a learnable projection matrix, and $\mathbf{v}_t$ is a learnable

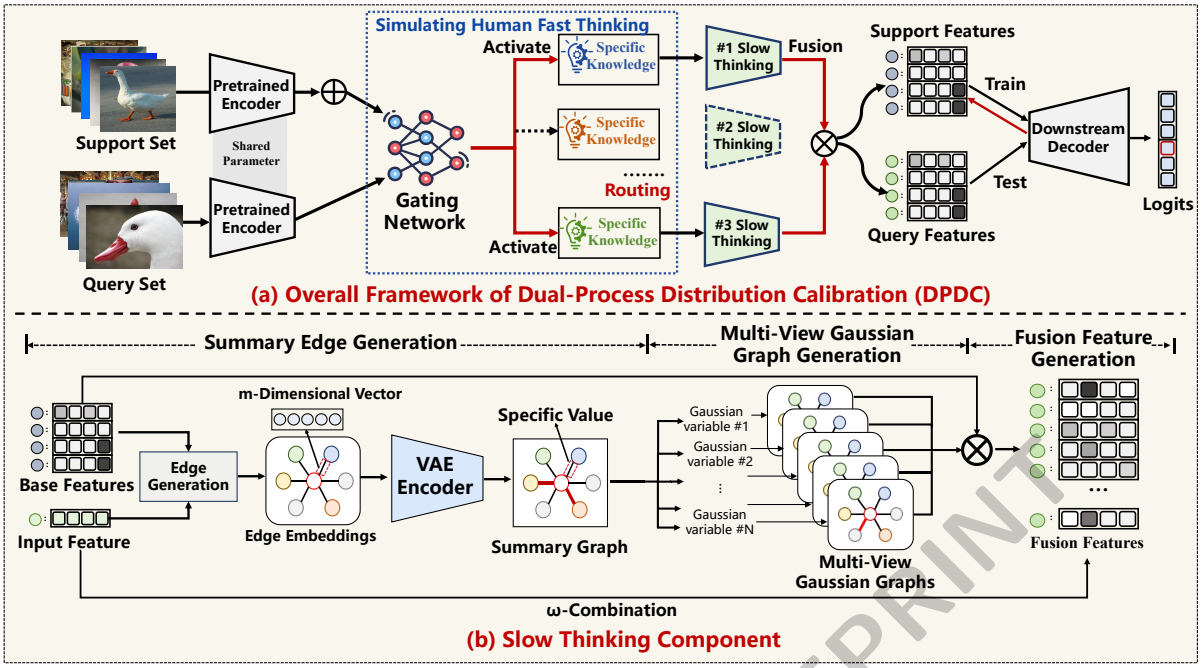

Figure 1: The overall architecture of the proposed Dual-Process Distribution Calibration (DPDC) framework for few-shot learning.

gating vector for the t-th slow-thinking component. The score g_t reflects the matching degree between the overall features of the current few-shot task and the knowledge within the t-th slow-thinking component.

To simulate the selective allocation of deep cognitive resources in fast thinking, we adopt a sparse gating mechanism, activating only the K ($K < T$) slow-thinking components most relevant to the current task for deep reasoning. Specifically, the gating network selects the top-K slow-thinking components with the highest scores g_t. Let $\mathcal{S} = \mathrm{Top}\,K(g_1, \ldots, g_T)$ denote the set of indices of these K highest scores. The corresponding attention weight α_t for the t-th component is then given by:

$$\alpha_t = \begin{cases} \dfrac{\exp(g_t)}{\sum_{j\in\mathcal{S}} \exp(g_j)}, & \text{if } t \in \mathcal{S} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus, for unselected components $t \notin \mathcal{S}$, $\alpha_t = 0$; for selected components, the weights are normalized via softmax over the selected scores, satisfying $\sum_{t=1}^{T} \alpha_t = 1$ with exactly K non-zero entries. These weights determine the contribution of activated components in later fusion. Hence, the fast-thinking component dynamically routes computational resources to the most relevant K slow-thinking modules based on the task context, providing a focused entry for subsequent fine-grained distribution calibration.

3.2 Slow-Thinking Component: Relation Inference and Divergent Sampling

Corresponding to the deep, deliberate System 2 in human cognition, the slow-thinking component forms the core of the distribution calibration framework. Each activated component t receives input features $\mathbf{h}_i$ from the fast-thinking component and performs relational reasoning using its stored base-class prototypes $\mathcal{P}^{(t)} = \{\mathbf{p}_j^{(t)} \in \mathbb{R}^d\}_{j=1}^{N_b}$ (where N_b denotes the number of base classes). Through two sub-processes, **relation integration** and **relation divergence**, it generates multi-perspective calibrated features $\mathcal{F}_i^{(t)} = \{\mathbf{f}_i^{(t,v)} \in \mathbb{R}^d\}_{v=1}^{V}$ for each input, achieving multi-perspective, divergent distribution calibration.

Relation Integration and Summary Graph Construction
Given an input sample feature $\mathbf{h}_i \in \mathbb{R}^d$, the slow-thinking component establishes associations with all base-class prototypes $\mathbf{p}_j^{(t)}$. A linear encoder $\phi : \mathbb{R}^{2d} \to \mathbb{R}^{d_e}$ maps the concatenated feature pair to an **edge embedding**, representing their potential relationship:

$$\mathbf{e}_{ij} = \phi([\mathbf{h}_i \| \mathbf{p}_j^{(t)}]), \tag{3}$$

where $[\cdot\|\cdot]$ denotes vector concatenation and d_e is the edge embedding dimensionality.

Cognitive aggregation is key to human associative inference, obtaining stable relational cognition between categories. We thus aggregate edge embeddings into a single, stable **Summary Graph**. Inspired by cognitive science [Greene *et al.*, 2001], human relational thinking involves numerous unconscious perceptions, mathematically modeled as sampling from a binomial distribution $B(n, \lambda)$ as $n \to \infty, \lambda \to 0$. According to the De Moivre–Laplace theorem [Spr, 2008], this binomial distribution can be approximated by a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to circumvent the issues when $n \to \infty, \lambda \to 0$. (Proof in supplementary material.)

Consequently, we employ variational inference to learn the distribution parameters of the Gaussian approximation directly from the edge embeddings $\mathbf{e}_{ij}$ via a parameterized deep network. Specifically, an MLP-based network $\mathcal{R}_\theta : \mathbb{R}^{d_e} \to \mathbb{R}^2$ computes the mean and standard deviation:

$$\mu_{ij}, \sigma_{ij} = \mathcal{R}_\theta(\mathbf{e}_{ij}), \tag{4}$$

where Softplus activates σ_{ij} to ensure non-negativity. These parameters yield a stable association strength m_{ij} representing the aggregated relation between the input and base class j. According to existing work [Huang *et al.*, 2020], the following formula guarantees that the obtained m_{ij} reflects the aggregated relational evidence analogous to human unconscious perception:

$$m_{ij} = \frac{1 + 2\mu_{ij}(\sigma_{ij})^2 - \sqrt{1 + 4(\mu_{ij})^2(\sigma_{ij})^4}}{2}. \tag{5}$$

All m_{ij} form the summary graph $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{1 \times N_b}$. This graph provides a robust starting point for divergent sampling, simulating the stable relational judgment from deep human thinking.

Multi-Perspective Relation Divergent Sampling

The component simulates human thought divergence to explore various inter-class relationships, avoiding overfitting to one summary graph. We apply **Divergent Gaussian Sampling** to the summary graph $\mathbf{M}$, generating V distinct Gaussian relation graphs $\{\mathbf{A}^{(v)}\}_{v=1}^{V}$ that each represent a relational perspective.

First, we compute an intermediate variable $\tilde{\alpha}$ from $\mathbf{M}$, adding Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ for diversity:

$$\tilde{\alpha}_{ij} = \sqrt{m_{ij}(1 - m_{ij})} \cdot \epsilon + m_{ij}. \tag{6}$$

For the v-th perspective, we conduct an additional Gaussian sampling to generate each element of the Gaussian relation graph $\mathbf{A}^{(v)}$:

$$s_{ij}^{(v)} = \mu_{ij} \cdot \tilde{\alpha}_{ij} + \sigma_{ij} \cdot \sqrt{\tilde{\alpha}_{ij}} \cdot \epsilon_{ij}^{(v)}, \quad \epsilon_{ij}^{(v)} \sim \mathcal{N}(0, 1), \tag{7}$$

$$\alpha_{ij}^{(v)} = \zeta(\tilde{\alpha}_{ij}^{(v)}) = \zeta(s_{ij}^{(v)} \cdot \tilde{\alpha}_{ij}), \tag{8}$$

where $\zeta(\cdot)$ normalizes $\alpha_{ij}^{(v)}$ to the range $[-1, 1]$. The matrix $\mathbf{A}^{(v)} = [\alpha_{ij}^{(v)}]$ defines the **Gaussian Relation Graph** for the v-th perspective. This approach generates diverse yet plausible relational hypotheses. And it ensures these hypotheses stay aligned with the robust relations in the summary graph.

Multi-Perspective Fused Feature Generation and Distribution Calibration

We employ the generated Gaussian relation graphs $\mathbf{A}^{(v)}$ for **distribution calibration** on the input feature. Each graph weights and fuses the input feature with base-class prototypes, producing a calibrated feature representation:

$$\mathbf{f}_i^{(t,v)} = (1 - \omega) \cdot (\mathbf{A}^{(v)}\mathcal{P}^{(t)}) + \omega \cdot \mathbf{h}_i, \tag{9}$$

where $\omega \in [0, 1)$ balances the input feature and base-class prototypes, and $\mathbf{A}^{(v)}\mathcal{P}^{(t)}$ denotes the linear combination of prototypes via $\mathbf{A}^{(v)}$.

Thus, a single slow-thinking component t generates multi-perspective calibrated features $\mathcal{F}_i^{(t)} = \{\mathbf{f}_i^{(t,v)}\}_{v=1}^{V}$ for input $\mathbf{h}_i$. These features incorporate base-class knowledge from different angles, achieving **multi-perspective, divergent calibration** of the data distribution for few-shot classes.

3.3 Feature Integration and Model Optimization

Multi-Perspective Feature Integration

Each input sample $\mathbf{h}_i$ receives multi-perspective calibrated feature sets $\{\mathcal{F}_i^{(t)} \mid \alpha_t > 0\}$ from activated slow-thinking components. The framework fuses these features via its gating weights α_t, producing the unified multi-perspective feature representation $\mathcal{F}_i = \{\mathbf{f}_i^v\}_{v=1}^{V}$:

$$\mathbf{f}_i^v = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{f}_i^{(t,v)}. \tag{10}$$

During training, support set samples use all V fused features $\mathcal{F}_i^s$ to represent the calibrated class distribution, training a lightweight logistic regression classifier. During testing, for each sample in the query set, we randomly select one perspective $\mathbf{f}_j$ from its fused features, which is then fed into the trained classifier to predict its class label. This design enables support set samples to achieve effective distribution calibration via multi-perspective features, while query set samples perform efficient inference with a single-perspective feature.

Joint Optimization Objective

Human relational reasoning is constrained by both current task evidence (posterior) and existing common sense or default assumptions (prior). To ensure stable and rational relation graph generation in the slow-thinking component, we add a variational inference-based regularization term $\mathcal{L}_{\text{KL}}$. Specifically, we employ two independent parameterized networks to simulate the posterior and prior cognitive sources respectively:

- **Posterior Path**: It takes edge embeddings $\mathbf{e}_{ij}$ as input and outputs distribution parameters $(\mu_{ij}^{\text{post}}, \sigma_{ij}^{\text{post}})$ and association strength m_{ij}^{post}. This path captures task-specific relational evidence for multi-perspective relation graph generation.
- **Prior Path**: Structurally identical to the posterior path but with independent inputs and parameters, it outputs corresponding prior distribution parameters $(\mu_{ij}^{\text{prior}}, \sigma_{ij}^{\text{prior}})$ and association strength m_{ij}^{prior}. It encodes a generic relational hypothesis reflecting default or common-sense knowledge, independent of the current task context.

The regularization loss $\mathcal{L}_{\text{KL}}$ minimizes the discrepancy between posterior and prior distributions and includes two parts:

Gaussian Distribution KL Divergence ($\mathcal{L}_{\text{KL-G}}$)

This term constrains the distribution of the Gaussian proxy in relational reasoning. It computes the KL divergence between the posterior Gaussian distribution $\mathcal{N}(\mu_{ij}^{\text{post}}, (\sigma_{ij}^{\text{post}})^2)$ and the prior distribution $\mathcal{N}(\mu_{ij}^{\text{prior}}, (\sigma_{ij}^{\text{prior}})^2)$:

$$\mathcal{L}_{\text{KL-G}} = \sum_{i,j} D_{\text{KL}}(q_{ij} \| p_{ij}) = \frac{1}{2} \sum_{i,j} \left[2 \log \frac{\sigma_{ij}^{\text{prior}}}{\sigma_{ij}^{\text{post}}} + \frac{(\sigma_{ij}^{\text{post}})^2 + (\mu_{ij}^{\text{post}} - \mu_{ij}^{\text{prior}})^2}{(\sigma_{ij}^{\text{prior}})^2} - 1 \right]. \tag{11}$$

This term ensures that the relational uncertainty (σ^{post}) and tendency (μ^{post}) learned from the current task do not excessively deviate from a reasonable default range.

Binomial Distribution Divergence ($\mathcal{L}_{\text{KL-B}}$)
This term directly constrains the scalar m_{ij} representing the robust association strength. As described in Section 3.2.1, the derivation of m_{ij} originates from modeling numerous human unconscious perceptions as sampling from a binomial distribution $B(n, \lambda)$ with $n \to \infty, \lambda \to 0$. Therefore, constraining m_{ij} essentially constrains the probability parameter λ of the underlying binomial distribution sampling. Referring to existing work [Liu and Jia, 2023], we can approximate the divergence term between these binomial distributions using m_{ij} via the following formula:

$$\mathcal{L}_{\text{KL-B}} = D_{\text{KL}}(B(n, \lambda^{\text{post}}) \| B(n, \lambda^{\text{prior}})) \approx \sum_{i,j} \left[m_{ij}^{\text{prior}} - m_{ij}^{\text{post}} + m_{ij}^{\text{post}} \cdot \left(\log m_{ij}^{\text{post}} - \log m_{ij}^{\text{prior}} \right) \right]. \tag{12}$$

Mathematically, this formula corresponds to an effective approximation and upper bound of the binomial KL divergence. Its motivation is to encourage the robust relational cognition learned by the model (m^{post}) to align with the default commonsense relational prior (m^{prior}), thereby enhancing the generalization and robustness of relational inference.

The overall objective function of the model is the weighted sum of the above losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \gamma \cdot \mathcal{L}_{\text{KL}} = \mathcal{L}_{\text{CE}} + \gamma \cdot (\mathcal{L}_{\text{KL-G}} + \mathcal{L}_{\text{KL-B}}), \tag{13}$$

where $\gamma \in (0, 1]$ is a hyperparameter. By optimizing the entire framework end-to-end via gradient descent, the model collaboratively learns how to dynamically screen relevant knowledge and perform deep, robust multi-perspective relational reasoning, ultimately achieving efficient and interpretable few-shot distribution calibration and classification.

4 Experiments

We evaluate DPDC through comprehensive experiments on three research questions:

- **RQ1 (Performance)**: How does DPDC compare with mainstream few-shot learning methods on standard and medical image datasets?
- **RQ2 (Mechanism)**: How do the fast-thinking and slow-thinking components (dynamic knowledge allocation and multi-perspective relation inference) and internal designs (e.g., divergent perspectives count) affect performance?
- **RQ3 (Interpretability)**: Do slow-thinking relation graphs correspond to semantic associations and provide credible explanations for model decisions?

We design three experimental parts: **Comparative Experiments** to compare DPDC with baseline methods across datasets; **Ablation Studies** to analyze the impact of core components and parameters; and **Visualization Analysis** to visualize inferred inter-class relations and verify interpretability. All experiments use the same setup for fair comparison.

4.1 Comparative Experiments

To comprehensively evaluate DPDC, we conduct systematic comparisons on three datasets: the general image dataset *mini*Imagenet [Russakovsky *et al.*, 2015], the fine-grained bird dataset CUB-200-2011 [Wah *et al.*, 2011], and the dermatology dataset Derm[1]. For *mini*Imagenet and CUB, we use the class splits from existing literature [Zhou *et al.*, 2025b]. For Derm, we split training, validation, and test sets by class in a 6:2:2 ratio. All methods use the same backbone (ResNet-18 pre-trained on each training set) for feature extraction. Experiments follow the full-way k-shot protocol on *mini*Imagenet and CUB. On Derm, we test 5-way and 20-way settings to simulate varying medical diagnostic complexity, repeating experiments for statistical stability. Selected baseline methods cover both classical and recent representative works, including: Prototypical Networks (PN) [Snell *et al.*, 2017], Matching Networks (MN) [Vinyals *et al.*, 2016], DC [Yang *et al.*, 2021], C2Net [Alam *et al.*, 2025], MF-SIN [Liu *et al.*, 2024], UNEM [Zhou *et al.*, 2025b], and CDN4 [Li *et al.*, 2025].

Results Analysis

Tables 1 and 2 present the classification accuracy of DPDC compared to baseline models. DPDC attains the highest accuracy across all settings, demonstrating the efficacy of simulating human dual-process cognition for distribution calibration from sparse data.

On the general dataset *mini*Imagenet, DPDC reaches 75.36% accuracy in the 1-shot task, exceeding the second-best model DC (66.58%) by 8.78%. The performance advantage persists as k increases, demonstrating consistent adaptability and robustness. This suggests that both the dynamic knowledge activation in fast-thinking and the multi-perspective relation divergence in slow-thinking improve the capacity to capture and generalize distribution patterns from few samples. DPDC also excels on the fine-grained CUB dataset.

On the medical dataset Derm, DPDC also exhibits strong generalization capability. Its accuracy significantly exceeds all baselines in both 5-way and 20-way tasks. Notably, in the 20-way 1-shot setting, DPDC reaches 39.49% accuracy, surpassing the best baseline by about 6%. These results confirm that the framework effectively addresses few-shot problems not only in general visual classification but also in specialized domains with substantial intra-class variation and complex

[1] www.dermnet.com

Method	*mini*Imagenet (20-way)					CUB (50-way)				
	1-shot	2-shot	4-shot	8-shot	16-shot	1-shot	2-shot	4-shot	8-shot	16-shot
PN 2017	40.39±0.64	47.96±0.56	54.61±0.52	59.85±0.51	62.21±0.46	35.09±0.60	42.90±0.47	50.65±0.40	54.94±0.49	56.26±0.44
MN 2016	54.43±0.75	65.89±0.63	71.29±0.46	77.97±0.36	81.19±0.41	44.01±0.54	50.68±0.52	55.12±0.41	60.53±0.49	63.79±0.39
DC 2021	66.58±1.52	70.60±1.12	80.40±0.94	83.18±0.81	84.62±0.70	31.39±0.87	45.03±0.98	60.23±1.06	72.37±0.84	77.81±0.75
C2Net 2025	-	-	-	-	-	33.11±0.12	42.47±0.09	49.91±0.08	54.69±0.07	57.28±0.07
MFSIN 2024	60.49±0.56	74.60±0.76	82.92±0.59	87.14±0.49	92.77±0.73	44.70±0.38	56.84±0.51	60.97±0.36	66.50±0.84	69.13±0.41
UNEM 2025b	60.01±0.63	67.04±0.55	73.83±0.53	79.05±0.54	85.02±0.48	41.49±0.66	48.21±0.35	63.90±0.49	67.74±0.35	69.69±0.42
CDN4 2025	64.40±0.82	72.44±0.81	79.30±0.61	85.98±0.58	88.44±0.58	34.70±0.19	46.90±0.17	56.51±0.15	63.05±0.14	67.03±0.13
DPDC (ours)	**75.36**±0.59	**87.86**±0.35	**92.91**±0.26	**95.53**±0.19	**96.22**±0.18	**48.30**±0.58	**61.60**±0.54	**68.93**±0.41	**74.52**±0.31	**78.95**±0.34

Table 1: Comparison of DPDC and baselines in accuracy on *mini*Imagenet and CUB datasets under few-shot scenarios. (Values after ± indicate the 95% confidence interval. **Bold** values indicate the best performance, while underlined values indicate the second-best performance.)

Method	Derm (5-way)					Derm (20-way)				
	1-shot	2-shot	4-shot	8-shot	16-shot	1-shot	2-shot	4-shot	8-shot	16-shot
PN 2017	50.20±3.04	59.90±3.41	69.90±2.81	71.70±2.69	62.21±0.46	32.52±1.69	39.68±1.63	41.78±1.36	45.88±1.58	56.26±0.44
MN 2016	57.44±2.23	64.52±2.17	70.72±2.28	75.92±1.97	77.46±1.87	33.92±1.10	42.46±1.03	49.84±0.92	53.04±1.14	60.92±0.99
DC 2021	53.99±1.77	56.98±1.50	73.12±1.34	78.72±1.46	80.90±1.32	31.42±1.35	45.02±1.29	57.84±1.82	66.06±1.51	67.11±1.46
MFSIN 2024	57.70±3.69	61.10±2.90	71.30±2.65	76.20±2.54	73.70±3.32	32.67±1.63	39.45±1.61	47.67±1.57	53.25±1.46	60.97±1.63
UNEM 2025b	59.30±2.84	63.80±3.74	68.50±2.95	69.00±2.69	85.02±0.48	32.37±1.68	39.50±1.64	42.12±1.36	46.60±1.59	69.69±0.42
CDN4 2025	52.10±3.81	57.10±3.13	66.90±2.58	67.40±3.27	66.70±2.65	32.43±1.65	41.12±1.67	44.68±1.40	50.05±1.37	58.75±1.53
DPDC (ours)	**63.90**±3.46	**74.68**±1.70	**81.02**±2.31	**86.78**±1.80	**86.90**±1.59	**39.49**±1.65	**52.02**±1.79	**62.74**±2.02	**71.07**±1.26	**72.06**±1.45

Table 2: Comparison of DPDC and baselines in accuracy on Derm dataset under few-shot scenarios.

semantics, validating the cognition-inspired mechanism's potential to enhance generalization and interpretability.

In summary, DPDC delivers stable, excellent performance across datasets and settings, with notable gains in low-sample scenarios, highlighting the effectiveness of mimicking human dual-process cognition for distribution calibration.

4.2 Ablation Studies

We perform ablation studies on *mini*Imagenet dataset to analyze the contributions of core DPDC components, focusing on: (1) the effect of the thought divergence mechanism in the slow-thinking component, and (2) the contribution of both fast and slow thinking components to overall performance.

Impact of Thought Divergence Degree

The slow-thinking component employs divergent Gaussian sampling to generate multi-perspective relation graphs for distribution calibration. Table **??** shows 5-way performance as perspectives V varies from 1 to 1000. Results confirm thought divergence is crucial for few-shot learning. Without divergence ($V = 1$), the accuracy is only 18.20%, indicating that a single relational view fails to capture inter-class association diversity in sparse data. Increasing V significantly improves performance; for 1-shot tasks, accuracy rises from 18.20% to 92.17% as V increases from 1 to 500. This effect is especially pronounced with very few support samples (1, 2-shot), where extensive divergence compensates for initial information scarcity by exploring numerous potential association hypotheses, enabling more robust category distribution inference.

However, with larger support sets (k), excessive divergence yields diminishing returns. For 16-shot tasks, performance peaks at 98.97% with $V = 20$, then slightly declines to 98.10% at $V = 1000$. This suggests that with more samples, relation aggregation captures true inter-class relationships more reliably, and excessive divergence may introduce noise. Notably, even at larger sample sizes (e.g., 8-shot, 16-shot), increasing V to 500 or 1000 does not severely degrade performance, demonstrating the summary-graph-based sampling effectively constrains graphs within a reasonable range. Considering computational efficiency and performance, $V = 100$ achieves near-optimal results in most settings, balancing accuracy and cost.

Contribution Analysis of Core Modules

Perspectives	miniImagenet (5-way)				
	1-shot	2-shot	4-shot	8-shot	16-shot
1	18.20	20.73	20.53	22.85	36.35
5	19.80	25.67	49.15	88.48	97.50
10	25.15	48.88	86.60	95.72	**98.75**
50	77.25	94.67	96.60	**98.72**	98.69
100	89.27	95.22	96.83	97.85	98.35
500	**92.17**	**96.10**	**97.92**	97.66	98.16
1000	92.05	95.75	96.72	97.12	98.10

Table 3: Performance of DPDC Model with Different Slow-Thinking Divergent Perspectives on miniImagenet Dataset (5-way)

To evaluate the necessity of the fast and slow thinking modules, we construct three variants for comparison:

- ***w/o* Fast Thinking**: Removes the fast-thinking component, disabling gated prototype matching and dynamic knowledge allocation.
- ***w/o* Slow Thinking**: Removes the slow-thinking component, omitting relation integration and divergent sam-

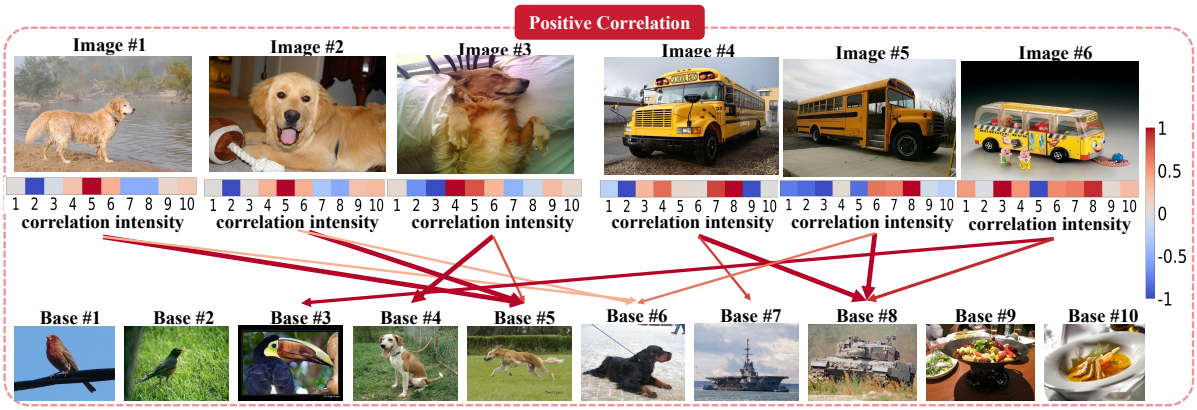

Figure 2: Visualization of relation graphs on the *mini*Imagenet dataset. Each target image (upper six images) points to the two base classes with the strongest positive correlation (lower ten images). Colors indicate correlation strength (red for positive, blue for negative).

variant	miniImagenet (20-way)				
	1-shot	2-shot	4-shot	8-shot	16-shot
w/o Fast and Slow Thinking	64.54$_{\pm 0.81}$	73.05$_{\pm 0.58}$	75.12$_{\pm 0.44}$	77.91$_{\pm 0.36}$	80.03$_{\pm 0.34}$
w/o Slow Thinking	66.57$_{\pm 0.78}$	79.16$_{\pm 0.82}$	82.12$_{\pm 0.45}$	83.50$_{\pm 0.35}$	86.45$_{\pm 0.36}$
w/o Fast Thinking	72.83$_{\pm 0.67}$	84.92$_{\pm 0.41}$	90.18$_{\pm 0.33}$	92.43$_{\pm 0.26}$	93.77$_{\pm 0.21}$
DPDC(Ours)	**75.36**$_{\pm 0.59}$	**87.86**$_{\pm 0.35}$	**92.91**$_{\pm 0.26}$	**95.53**$_{\pm 0.19}$	**96.22**$_{\pm 0.18}$

Table 4: Performance comparison of DPDC and its variants on the miniImagenet dataset. (Values after $\pm$ indicate the 95% confidence interval.)

pling. It directly uses inter-class edge embeddings as the association graph to combine with base-class prototypes.

- ***w/o* Fast and Slow Thinking**: This variant removes both components.

Table **??** compares the performance of the complete DPDC framework with its variants on 20-way tasks. The full model achieves the best accuracy for all k, confirming the dual-process collaborative design's effectiveness. Removing the slow-thinking component causes a marked performance drop, especially in low-sample scenarios, indicating that fast matching alone cannot achieve precise distribution calibration; deep relational reasoning and multi-perspective divergence are essential for robustness and accuracy. Removing the fast-thinking component also degrades performance, albeit less severely, demonstrating its role in focusing computational resources. The variant lacking both components performs worst, further validating both modules.

In summary, the ablation studies show that: (1) The thought divergence mechanism in slow thinking is central to DPDC, enhancing distribution calibration by exploring multiple association hypotheses in few-shot scenarios, while moderate divergence ($V = 100$) suffices when samples are abundant; (2) Both components contribute significantly to final performance, with the deep relation integration and divergence in slow thinking exerting a stronger influence.

4.3 Visualization Analysis

We visualize Gaussian relation graphs from the slow-thinking component to examine whether DPDC constructs inter-class relationships aligned with human cognitive patterns. We manually select ten base classes (three birds, three dogs, two vehicles, and two food items) and train corresponding slow-thinking components, then visualize a Gaussian association graph for each test image.

Figure 2 displays the results, with two target sets: three dog images (left) and three school bus images (right). Each connects via arrows to the two base classes with highest positive correlation. For dog targets, the most relevant base classes consistently belong to the dog categories (Base Classes 4, 5, and 6), indicating correct semantic association. For school bus targets, Base Class 8 (tank) consistently appears among the strongest positive correlations, reflecting visual similarity between school buses and tanks. These results demonstrate that the slow-thinking component generates semantically meaningful inter-class associations aligned with human intuition, validating DPDC's effectiveness in interpretable, human-like relational reasoning.

5 Conclusion

Inspired by human dual-process cognition, this paper proposes a Dual-Process Distribution Calibration framework (DPDC) for few-shot image classification. DPDC systematically implements a fast-slow thinking mechanism. A gated prototype matching component rapidly screens and allocates prior knowledge, whereas a variational inference component with divergent sampling calibrates inter-class relationships from multiple relational perspectives. Experiments on several datasets demonstrate that DPDC consistently surpasses state-of-the-art methods. The resulting relation graphs exhibit human-aligned semantic interpretability and offer transparent justification for model decisions. This work provides new insights for constructing efficient and trustworthy human-like learning systems. Future research will explore alternative computational realizations of the dual-process cognitive framework and extend its applicability to multimodal learning and more complex decision-making tasks.

Ethics Statement

This work uses publicly available benchmark datasets, including the Derm dermatology image dataset. The version of Derm used in this study does not contain patient-identifiable information, such as names, patient IDs, addresses, or other metadata that could directly identify individuals. The authors declare no competing interests. The code for DPDC is available at https://github.com/YuchenLiu1225/DPDC-IJCAI26.

Acknowledgments

This work was supported by the Science and Technology Innovation Program of Hunan Province (No.2025ZYC007) and the Innovation Funding of the Institute of Computing Technology, Chinese Academy of Sciences (No.E561060).

References

[Alam *et al.*, 2025] Nur Alam, Md Rakibul Islam, L Minh Dang, Mariya Kibtiya, Adnan Hossan Rifat, and Moon Hyeonjoon. C2-net: Improving feature extraction and alignment for few-shot fine-grained image classification. *Research Square*, April 2025. Preprint.

[Finn *et al.*, 2017] 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.

[Greene *et al.*, 2001] Anthony J. Greene, Barbara A. Spellman, Jeffery A. Dusek, Howard B. Eichenbaum, and William B. Levy. Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. *Memory & Cognition*, 29(6):893–902, 2001.

[Huang *et al.*, 2020] Hengguan Huang, Fuzhao Xue, Hao Wang, and Ye Wang. Deep graph random process for relational-thinking-based speech recognition. In *International Conference on Machine Learning*, pages 4531–4541. PMLR, 2020.

[Kahneman, 2011] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[Li *et al.*, 2025] Xiaoxu Li, Shuo Ding, Jiyang Xie, Xiaochen Yang, Zhanyu Ma, and Jing-Hao Xue. Cdn4: A cross-view deep nearest neighbor neural network for fine-grained few-shot classification. *Pattern Recognition*, 163:111466, 2025.

[Liu and Jia, 2023] Yuchen Liu and Ziyu Jia. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*, 2023.

[Liu *et al.*, 2024] Shudong Liu, Wenlong Zhong, Furong Guo, Jia Cong, and Boyu Gu. Fine-grained few-shot image classification based on feature dual reconstruction. *Electronics*, 13(14), 2024.

[Pennycook *et al.*, 2015] Gordon Pennycook, Jonathan A. Fugelsang, and Derek J. Koehler. What makes us think? a three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80:34–72, 2015.

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[Spr, 2008] *De Moivre–Laplace Theorem*, pages 356–358. Springer New York, New York, NY, 2008.

[Sun *et al.*, 2025] Zhe Sun, Mingyang Wang, Xiangchen Ran, and Pengfei Guo. Feature reconstruction-guided transductive few-shot learning with distribution statistics optimization. *Expert Systems with Applications*, 270:126555, 2025.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *california institute of technology*, 2011.

[Wang *et al.*, 2020] Yaqing Wang, Quanming Yao, James Tin Yau Kwok, and Lionel Ming-Shuan Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, June 2020.

[Wang *et al.*, 2024] Xuan Wang, Zhong Ji, Yanwei Pang, and Yunlong Yu. A cognition-driven framework for few-shot class-incremental learning. *Neurocomputing*, 600:128118, 2024.

[Wu *et al.*, 2025] Yiru Wu, Yuhang Xia, Hao Li, Lixin Yuan, Junyang Chen, Jun Liu, Tong Lu, and Shaohua Wan. Deconfound semantic shift and incompleteness in incremental few-shot semantic segmentation. In *AAAI Conference on Artificial Intelligence*, 2025.

[Yang *et al.*, 2021] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021.

[Zhou *et al.*, 2025a] Chunpeng Zhou, Zhi Yu, Xilu Yuan, Sheng Zhou, Jiajun Bu, and Haishuai Wang. Less is more: A closer look at semantic-based few-shot learning. *Information Fusion*, 114:102672, 2025.

[Zhou *et al.*, 2025b] Long Zhou, Fereshteh Shakeri, Aymen Sadraoui, Mounir Kaaniche, Jean-Christophe Pesquet, and Ismail Ben Ayed. UNEM: UNrolled Generalized EM for Transductive Few-Shot Learning. In *CVPR*, June 2025.