DESCRIPTION-AND-SUPPORT GUIDED PROTOTYPE FUSION FOR FEW-SHOT IMAGE CLASSIFICATION

Yuxiang Cheng *Tongyue Sun* *Kian Ming Lim** *Chin Poo Lee**

University of Nottingham Ningbo China, Ningbo, China

ABSTRACT

Few-shot classification benefits from frozen CLIP visual features, yet direct support-query matching underuses category-level descriptions and generated supports. We propose Description-and-Support Guided Prototype Fusion (DSGPF), which uses frozen CLIP visual matching and adds a trainable prototype branch for episode-level prediction. From category names, DSGPF derives structured visual descriptions, encodes them as semantic embeddings, and formats them as prompts for class-conditioned generated supports. During each episode, the Description-and-Support Guided Prototype Branch (DSGPB) constructs a visual prototype from real and generated supports and a description-conditioned prototype from real supports and semantic embeddings. Calibrated Probability Fusion (CPF) resolves posterior confidence-scale mismatch between this prototype branch and frozen visual matching through validation-selected calibration. Experiments on CIFAR-FS, FC100, and miniImageNet show consistent gains over frozen visual matching in 5-way 1-shot and 5-way 5-shot evaluation. Ablations show that descriptions and generated supports strengthen the prototype branch, and fusion analysis confirms that CPF outperforms direct posterior fusion rules.

***Index Terms*—** Few-shot classification, vision-language models, prototype fusion, generated supports

1. INTRODUCTION

Few-shot classification recognizes novel categories from a few labeled support images. Metric-based methods address this setting by comparing queries with support examples or class prototypes [1, 2, 3]. Recent studies show that representation quality and evaluation protocol strongly affect few-shot performance [4, 5], motivating prediction rules that combine transferable features with episode-specific support evidence. These findings point to two practical requirements: transferable representations and episode-adaptive prediction. A useful predictor combines reliable visual matching with class-specific evidence shaped by the sampled support set.

Vision-language foundation models provide strong transferable representations. CLIP aligns image and text features through large-scale contrastive pretraining [6], and has motivated prompt learning, adapters, and support-feature retrieval for few-shot transfer [7, 8, 9, 10]. These methods demonstrate effective CLIP adaptation from limited support evidence. Description cues and generated images introduce heterogeneous class-side evidence that needs episode-level organization before prediction. Language models and text-to-image generators derive visual descriptions and generated supports from category names [11, 12, 13, 14]. The resulting design problem is how to integrate class-side semantic and generated visual evidence with strong episode-level visual matching. We frame integration at the episode level, where semantic descriptions, generated images, and real supports are mapped to the same five-way label space before prediction. A shared episode label space allows semantic, generated, and real-support cues to form comparable prototype posteriors before fusion.

We use frozen CLIP visual matching as the reference visual branch. The CLIP image encoder remains frozen, and support-query matching in CLIP visual space serves as the Visual Matching Branch (VMB). Structured descriptions and generated supports are converted into an episode-level prototype posterior for the sampled classes. This places semantic and generated cues where few-shot supervision is naturally expressed: prototype construction conditioned on the current support set.

We propose Description-and-Support Guided Prototype Fusion (DSGPF). First, Class-Level Semantic Decomposition (CLSD) turns category names into structured visual-discriminative descriptions, which are encoded as semantic embeddings and used to guide class-conditioned generated supports. Second, the Description-and-Support Guided Prototype Branch (DSGPB) constructs prototypes from real supports, generated supports, and description-conditioned support features. Third, Calibrated Probability Fusion (CPF) aligns and combines the DSGPB posterior with the frozen VMB posterior using validation-selected scalar parameters.

Our contributions are summarized as follows:

- We propose DSGPF, an episode-level semantic-visual prototype fusion framework that uses frozen CLIP visual matching.
- We introduce CLSD and DSGPB to convert category names, structured descriptions, and generated supports into complementary prototype evidence.
- Experiments on CIFAR-FS, FC100, and miniImageNet show consistent gains over frozen visual matching, with ablations validating structured descriptions, generated supports, and calibrated fusion.

2. RELATED WORK

Few-shot metric and prototype learning. Few-shot classification is commonly formulated as episodic support-query prediction. Matching Networks, Prototypical Networks, and Relation Networks establish attention-based matching, prototype averaging, and learned comparison functions [1, 2, 3]. Later methods improve metric learning through task-dependent scaling, differentiable convex optimization, strong supervised embeddings, set-to-set adaptation, and dense region matching [15, 16, 17, 18, 19]. These results support a practical view of few-shot recognition: a strong representation and a well-calibrated episode classifier are both central to performance. DSGPF follows the prototype-learning view and enriches episode prototypes with structured descriptions and generated supports.

*Corresponding authors.

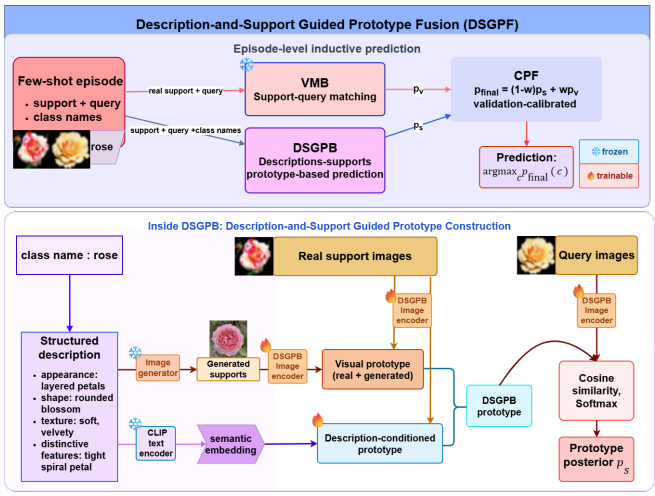


Fig. 1. Overview of DSGPF and the prototype construction inside DSGPB. VMB performs frozen CLIP support-query matching to produce p_v, and DSGPB constructs a prototype posterior p_s from real supports, generated supports, and structured descriptions. CPF fuses the calibrated posteriors as $p_{\text{final}} = (1-w)p_s + wp_v$.

CLIP adaptation for few-shot recognition. CLIP provides transferable image-text representations and strong zero-shot recognition [6]. Few-shot adaptation methods improve CLIP with feature adapters, learned prompts, multi-modal prompts, task residuals, and support-feature retrieval [7, 8, 10, 9]. These approaches provide strong CLIP-based reference points for few-shot transfer. DSGPF uses frozen CLIP visual matching and calibrates it with a separate Description-and-Support Guided Prototype Branch.

Semantic descriptions and generated supports. Semantic priors and generated examples alleviate data scarcity in low-shot recognition [20]. Recent foundation-model methods use LLM descriptions, name-conditioned synthetic images, semantic-aided prototypes, and cascaded language-generation priors for few-shot transfer [11, 12, 13, 14, 21]. DSGPF uses these resources inside episode-specific prototype construction and calibrates the resulting posterior with frozen visual matching.

3. METHOD

DSGPF addresses 5-way few-shot classification by combining frozen CLIP visual matching with an episode-level prototype branch. As shown in figure 1, class names are first converted into structured visual descriptions, semantic embeddings, and generated supports. During each episode, DSGPB builds prototypes from real supports, generated supports, and description-conditioned support features; CPF then calibrates and fuses the DSGPB posterior with the frozen visual-matching posterior. The branch design makes the evidence sources explicit: VMB provides direct visual affinity, and DSGPB expresses semantic and generated cues as prototypes in the episode label space.

The CLIP image encoder used by VMB and the CLIP text encoder used by Structured Semantic Embedding (SSE) remain frozen. The language model used for CLSD and the text-to-image generator used for Class-Conditioned Synthetic Visual Generation (CSVG) are fixed during class-level resource preparation. The trainable components are the DSGPB image encoder, semantic projection layers, and description-guided modulation module, optimized on base-class episodes. The prototype mixture and CPF calibration parameters are selected on validation episodes and fixed for novel-class evaluation.

3.1. Problem Setup and Protocol

We consider 5-way K-shot classification with $K \in \{1, 5\}$. An episode contains a labeled support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{5K}$, an unlabeled query set $\mathcal{Q} = \{x_j^q\}$, and five episode classes $\mathcal{C} = \{c_1, \ldots, c_5\}$. Each class c has a category name n_c; WordNet identifiers or folder names are mapped to readable category names before description construction. Each query image is classified into one of the five episode classes. Class-level resources are built from category names before episodic evaluation. Validation and novel images enter DSGPF through the labeled support examples of sampled episodes.

3.2. Structured Description and Generated Support Construction

For each class c, DSGPF prepares a structured visual description d_c, a semantic embedding e_c, and a set of class-conditioned generated supports $\tilde{\mathcal{S}}_c$. The category name n_c is the language-side input, and the resulting resources are reused across all reported episodes involving that class.

CLSD. CLSD produces d_c from n_c with a fixed reasoning-guided template. The template asks for visual-discriminative fields covering appearance, shape, texture, part structure, context, distinctive cues, likely confusions, and a final summary. Compared with compact, sometimes ambiguous category names, structured descriptions provide stable visual cues for prototype construction.

SSE. SSE maps each structured description into a semantic vector using a frozen CLIP text encoder. Let f_{text} denote the text encoder with its projection head. For each class c,

$$e_c = \frac{f_{\text{text}}(d_c)}{\|f_{\text{text}}(d_c)\|_2}, \quad e_c \in \mathbb{R}^{d_t}. \tag{1}$$

The normalized embedding guides description-conditioned support encoding for the classes sampled in the current episode.

CSVG. CSVG uses d_c with a frozen text-to-image generator to produce fixed generated supports $\tilde{\mathcal{S}}_c = \{\tilde{x}_{c,m}\}_{m=1}^{M_c}$. Generated supports enter DSGPB as class-conditioned visual evidence for prototype construction; the main experiments use them directly with the same image preprocessing as real supports. Generated supports enter the prediction path through class prototypes in the DSGPB feature space, and the description-conditioned path uses real support images as anchors whose features are modulated by class semantics.

3.3. DSGPB Prototype Branch

DSGPB produces logits $\ell_s(c \mid x^q)$ by constructing episode-specific prototypes from real supports, generated supports, and structured descriptions. Let $\phi_\theta(\cdot)$ denote the DSGPB image encoder, which maps each image to a d-dimensional feature.

Real-generated visual prototype. For class c, let $\mathcal{S}_c = \{x_i : (x_i, y_i) \in \mathcal{S}, y_i = c\}$ be its real supports and let $\tilde{\mathcal{S}}_c^{(K)}$ be its generated supports in the K-shot setting. DSGPB computes

$$r_c = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} \phi_\theta(x), \quad g_c = \frac{1}{|\tilde{\mathcal{S}}_c^{(K)}|} \sum_{\tilde{x} \in \tilde{\mathcal{S}}_c^{(K)}} \phi_\theta(\tilde{x}). \tag{2}$$

The real and generated prototypes are averaged with their sample counts:

$$b_c = \frac{|\mathcal{S}_c| r_c + |\tilde{\mathcal{S}}_c^{(K)}| g_c}{|\mathcal{S}_c| + |\tilde{\mathcal{S}}_c^{(K)}|}. \tag{3}$$

For the No Gen. ablation, $b_c = r_c$. In the main experiments, $|\tilde{\mathcal{S}}_c^{(K)}| = K$; generated-support ablations remove them, replace them with mismatched-class samples, or keep the class-matched CSVG output.

Description-guided support encoding. DSGPB also constructs a description-conditioned prototype from real supports and e_c. The semantic embedding is projected into the DSGPB feature space:

$$q_c^{(1)} = W_1 e_c, \quad q_c^{(2)} = W_2 e_c, \tag{4}$$

where W_1 and W_2 are learnable matrices. The projected vectors provide lightweight spatial and channel conditioning inside the image encoder: the channel path emphasizes description-relevant dimensions, and the spatial path injects the class cue into intermediate feature maps. This yields $\phi_\theta^{\text{desc}}(x, e_c)$. DSGPB then aggregates

$$u_c = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} \phi_\theta^{\text{desc}}(x, e_c). \tag{5}$$

Generated supports contribute through b_c, and structured descriptions modulate real support features through u_c. During scoring, query images are encoded once with the unconditioned DSGPB encoder ϕ_θ.

Guided prototype and DSGPB logits. The final DSGPB prototype mixes the real-generated visual prototype and the description-conditioned prototype:

$$a_c = \rho b_c + (1 - \rho) u_c, \tag{6}$$

where $\rho \in [0, 1]$ is selected on validation episodes after DSGPB training. Given a query image x^q, DSGPB computes cosine logits

$$\ell_s(c \mid x^q) = \left\langle \frac{\phi_\theta(x^q)}{\|\phi_\theta(x^q)\|_2}, \frac{a_c}{\|a_c\|_2} \right\rangle. \tag{7}$$

During DSGPB training, these logits are divided by a fixed temperature $t = 0.2$; at inference time, CPF calibrates the raw logits with the validation-selected temperature T_s.

3.4. VMB

VMB is the frozen visual matching branch. It uses a frozen CLIP image encoder f_{img} and real support/query images. Each image is encoded as $\hat{v}(x) = f_{\text{img}}(x)/\|f_{\text{img}}(x)\|_2$. For query x^q, VMB computes support affinities

$$A_i(x^q) = \frac{\exp(\beta \langle \hat{v}(x^q), \hat{v}(x_i) \rangle)}{\sum_{(x_j, y_j) \in \mathcal{S}} \exp(\beta \langle \hat{v}(x^q), \hat{v}(x_j) \rangle)}, \tag{8}$$

where β is a fixed concentration parameter. The class probability mass is aggregated over support labels, and the log mass gives VMB logits:

$$\ell_v(c \mid x^q) = \log \left(\sum_{(x_i, y_i) \in \mathcal{S} : y_i = c} A_i(x^q) + \epsilon \right). \tag{9}$$

3.5. CPF

DSGPB and VMB produce posteriors with different confidence scales. CPF applies post-hoc temperature scaling to both branches,

$$p_s = \text{softmax}(\ell_s / T_s), \quad p_v = \text{softmax}(\ell_v / T_v), \tag{10}$$

and fuses the two episode-level distributions as

$$p_{\text{final}} = (1 - w) p_s + w p_v. \tag{11}$$

The temperatures T_s, T_v and fusion coefficient w are selected from predefined candidate sets on validation episodes and then fixed for novel-class evaluation. CPF is applied after DSGPB training: validation episodes select fixed temperatures and a fusion weight, then novel episodes reuse those scalars for branch-level posterior fusion.

3.6. Training Objective and Optimization

DSGPF is trained on base classes in two stages. We first train the DSGPB image encoder on base-class images, then train DSGPB episodically using prepared semantic embeddings and generated supports. During episodic training, the projection layers W_1, W_2, modulation module, and DSGPB image encoder are optimized with

$$\mathcal{L}_{\text{DSGPB}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{tri}} + \lambda_{\text{bpa}} \mathcal{L}_{\text{bpa}}. \tag{12}$$

Here $\mathcal{L}_{\text{cls}}$ is the episodic cross-entropy loss on the description-conditioned comparison space. The auxiliary term $\mathcal{L}_{\text{tri}}$ encourages local consistency among projected description features, description-conditioned support features, and same-class generated-support features. The alignment term $\mathcal{L}_{\text{bpa}}$ matches real-support and description-conditioned prototypes through a bidirectional contrastive objective. In the main runs, $\lambda_{\text{bpa}} = 0.1$. The CLIP text encoder, CLIP image encoder, and text-to-image generator remain frozen throughout training and evaluation.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets and episodes. We evaluate DSGPF on CIFAR-FS, FC100, and miniImageNet using the standard training/validation/test splits: 64/16/20 classes for CIFAR-FS and miniImageNet, and 60/20/20 classes for FC100. We report 5-way 1-shot and 5-way 5-shot novel-class accuracy over 2,000 test episodes, each with 15 query images per class, as mean accuracy with 95% confidence intervals.

Protocol and implementation. All experiments follow the protocol in section 3.1. Each query is predicted from the current support set and prepared class resources; CPF parameters (T_s, T_v, w) are selected on validation episodes and fixed for novel-class evaluation. DSGPB uses Visformer-Tiny with 224×224 inputs and 384-dimensional features. VMB uses a frozen OpenAI CLIP ViT-B/32 image encoder, and SSE uses the corresponding frozen CLIP text encoder. CLSD uses Claude Sonnet with a fixed structured template, CSVG uses Qwen-Image 2, and DSGPF uses K generated supports per class. VMB uses $\beta = 20$, and CPF searches $T_s, T_v \in \{0.5, 0.7, 1.0, 1.5, 2.0\}$ and $w \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Training uses supervised base-class pretraining followed by episodic DSGPB optimization with AdamW.

4.2. Main Results: Comparison with Published State-of-the-Art Few-Shot Methods

Tables 1 and 2 compare DSGPF with recent few-shot classification methods on miniImageNet, CIFAR-FS, and FC100. DSGPF achieves strong results across all three benchmarks: 90.73%/97.01% on miniImageNet, 86.44%/93.40% on CIFAR-FS, and 59.65%/74.95% on FC100 for 1-shot/5-shot evaluation. These

results place DSGPF among the strongest listed methods across both fine-grained and coarse-grained few-shot benchmarks.

Table 1. Comparison with recent state-of-the-art methods on miniImageNet. Accuracy is 5-way (%); – denotes unavailable results.

Method	Venue	miniImageNet	
		1-shot	5-shot
FewTURE [22]	NeurIPS'22	68.02 ± 0.88	84.51 ± 0.53
CPEA [23]	ICCV'23	71.97 ± 0.65	87.06 ± 0.38
LPE-CLIP [24]	WACV'23	71.64 ± 0.40	79.67 ± 0.32
SP [25]	CVPR'23	72.31 ± 0.40	83.42 ± 0.30
FeatWalk [26]	AAAI'24	70.21 ± 0.44	87.38 ± 0.27
BECLR [27]	ICLR'24	75.74 ± 0.62	84.93 ± 0.33
SIFT [28]	IJCV'24	77.31 ± 0.67	86.95 ± 0.53
SemFew [13]	CVPR'24	78.94 ± 0.66	86.49 ± 0.50
KTPP [29]	ACM MM'24	76.71 ± 0.37	86.46 ± 0.27
UAP [30]	NeurIPS'24	81.63 ± 0.28	79.05 ± 0.19
ECER [31]	AAAI'25	81.14 ± 0.15	–
SAFF [32]	CVPRW'25	71.51 ± 0.63	87.19 ± 0.36
Tip-Adapter [9]	ECCV'22	73.02 ± 0.39	79.20 ± 0.32
DSGPF	Ours	$\mathbf{90.73 \pm 0.29}$	$\mathbf{97.01 \pm 0.14}$

Table 2. Comparison with recent state-of-the-art methods on CIFAR-FS and FC100. Accuracy is 5-way (%); – denotes unavailable results.

Method	Venue	CIFAR-FS		FC100	
		1-shot	5-shot	1-shot	5-shot
FewTURE [22]	NeurIPS'22	77.76 ± 0.81	88.90 ± 0.59	47.68 ± 0.78	63.81 ± 0.75
CPEA [23]	ICCV'23	–	–	47.24 ± 0.58	65.02 ± 0.60
LPE-CLIP [24]	WACV'23	80.62 ± 0.41	86.22 ± 0.33	–	–
SP [25]	CVPR'23	82.18 ± 0.40	88.24 ± 0.32	48.53 ± 0.38	61.55 ± 0.41
BECLR [27]	ICLR'24	–	–	45.21 ± 0.50	60.02 ± 0.43
AMMD [33]	PR'24	77.61 ± 0.46	88.08 ± 0.33	46.86 ± 0.43	63.79 ± 0.44
SemFew [13]	CVPR'24	84.34 ± 0.67	89.11 ± 0.54	54.27 ± 0.77	65.02 ± 0.72
KTPP [29]	ACM MM'24	80.80 ± 0.43	90.01 ± 0.29	–	–
ECER [31]	AAAI'25	86.01 ± 0.35	–	57.34 ± 0.31	–
SAFF [32]	CVPRW'25	78.48 ± 0.67	90.30 ± 0.41	47.17 ± 0.59	66.00 ± 0.57
Tip-Adapter [9]	ECCV'22	81.39 ± 0.38	86.46 ± 0.31	52.01 ± 0.40	59.30 ± 0.39
DSGPF	Ours	$\mathbf{86.44 \pm 0.34}$	$\mathbf{93.40 \pm 0.22}$	$\mathbf{59.65 \pm 0.48}$	$\mathbf{74.95 \pm 0.40}$

4.3. Strengthening Frozen Visual Matching through Prototype Fusion

We compare the frozen visual-matching baseline VMB, direct posterior averaging (VMB+DSGPB Avg.), and DSGPF with CPF. DSGPB constructs prototype evidence from real supports, generated supports, and description-conditioned support features. CPF calibrates the branch confidence scales before fusion.

Table 3. Controlled fusion results. VMB is the frozen visual-matching baseline; Avg. directly averages branch posteriors; DSGPF applies CPF.

Prediction rule	CIFAR-FS		FC100		miniImageNet	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
VMB only	75.36 ± 0.44	90.14 ± 0.26	54.42 ± 0.50	72.27 ± 0.42	88.31 ± 0.37	96.67 ± 0.15
VMB+DSGPB Avg.	80.09 ± 0.41	92.32 ± 0.24	56.45 ± 0.49	74.16 ± 0.40	89.17 ± 0.35	96.87 ± 0.14
DSGPF	$\mathbf{86.44 \pm 0.34}$	$\mathbf{93.40 \pm 0.22}$	$\mathbf{59.65 \pm 0.48}$	$\mathbf{74.95 \pm 0.40}$	$\mathbf{90.73 \pm 0.29}$	$\mathbf{97.01 \pm 0.14}$

Table 3 shows consistent gains from prototype fusion. Direct averaging improves over VMB in all six settings, and CPF further improves over averaging, with the largest margins in 1-shot evaluation.

4.4. Ablation of Structured Descriptions and Generated Supports

We isolate structured descriptions and generated supports with fixed first-stage visual initialization. Each variant retrains second-stage DSGPB and reports both prototype-branch accuracy and final fused accuracy. The variants compare category-name descriptions, CLSD descriptions, mismatched generated supports, and class-matched CSVG supports.

Table 4. Ablation of descriptions and generated supports. DSGPB reports prototype-branch accuracy; DSGPF reports final fused accuracy.

Description	Generated-support setting	Prediction rule	FC100		miniImageNet	
			1-shot	5-shot	1-shot	5-shot
Category name	No Gen.	DSGPB	43.30 ± 0.37	55.35 ± 0.38	61.16 ± 0.46	75.37 ± 0.35
Category name	No Gen.	DSGPF	56.66 ± 0.48	74.24 ± 0.40	88.81 ± 0.35	96.89 ± 0.14
CLSD	No Gen.	DSGPB	43.47 ± 0.39	55.07 ± 0.38	70.98 ± 0.39	76.71 ± 0.34
CLSD	No Gen.	DSGPF	56.78 ± 0.49	73.80 ± 0.40	90.42 ± 0.31	96.93 ± 0.14
CLSD	Mismatched Gen.	DSGPB	46.74 ± 0.40	57.23 ± 0.39	70.58 ± 0.38	76.71 ± 0.33
CLSD	Mismatched Gen.	DSGPF	57.88 ± 0.48	74.38 ± 0.40	90.41 ± 0.30	96.99 ± 0.14
CLSD	CSVG	DSGPB	48.52 ± 0.42	58.47 ± 0.40	74.29 ± 0.36	78.72 ± 0.31
CLSD	CSVG	DSGPF	$\mathbf{59.65 \pm 0.48}$	$\mathbf{74.95 \pm 0.40}$	$\mathbf{90.73 \pm 0.29}$	$\mathbf{97.01 \pm 0.14}$

Table 4 shows that CLSD strengthens DSGPB most clearly on miniImageNet. CSVG gives the strongest prototype-branch and fused results among the tested variants. Mismatched generated supports improve several settings relative to No Gen., suggesting a regularization effect from additional generated visual features; the CSVG gap indicates the value of class-conditioned generation. The CLSD-only comparison isolates the role of structured semantics in support conditioning, and the CSVG comparison shows that generated supports are most useful when their visual content is aligned with the episode class. These trends match the design of DSGPB, which converts each source into class-level prototype evidence before CPF combines branch posteriors.

4.5. Calibration of Prototype and Visual-Matching Predictions

We evaluate CPF against direct fusion rules applied to DSGPB and VMB: probability averaging, product-rule fusion, and max-probability ensembling.

Table 5. Calibration and fusion ablation. Results are accuracies in percent.

Prediction rule	CIFAR-FS		FC100		miniImageNet	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Avg.	80.09	92.32	56.45	74.16	89.17	96.87
Prod.	81.19	92.60	57.17	74.44	89.54	96.94
MaxP	76.71	90.94	54.65	72.51	88.43	96.69
CPF	**86.44**	**93.40**	**59.65**	**74.95**	**90.73**	**97.01**

Table 5 shows that CPF obtains the best accuracy across the compared fusion rules in all six settings. The result highlights the importance of validation-selected temperature scaling before combining prototype and visual matching predictions.

5. CONCLUSION

We presented DSGPF, a description-and-support guided prototype fusion framework for few-shot classification. DSGPF forms prototype posteriors from generated and description-conditioned supports, then calibrates them with frozen CLIP visual matching. Experiments show consistent gains across three benchmarks, and ablations validate the proposed components. These results support converting semantic and generated resources into calibrated episode-level prototype evidence. This episode-level framing ties foundation-model resources to the sampled support classes and uses visual matching as a calibrated reference.

6. REFERENCES

[1] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *Proc. NeurIPS*, 2016, vol. 29.

[2] Jake Snell, Kevin Swersky, and Richard S. Zemel, "Prototypical networks for few-shot learning," in *Proc. NeurIPS*, 2017, vol. 30.

[3] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. CVPR*, 2018, pp. 1199–1208.

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang, "A closer look at few-shot classification," in *Proc. ICLR*, 2019.

[5] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola, "Rethinking few-shot image classification: A good embedding is all you need?," in *Proc. ECCV*. Springer, 2020, pp. 266–282.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, vol. 139, pp. 8748–8763.

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao, "CLIP-adapter: Better vision-language models with feature adapters," *IJCV*, vol. 132, pp. 581–595, 2024.

[8] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *IJCV*, vol. 130, pp. 2337–2348, 2022.

[9] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li, "Tip-adapter: Training-free adaption of CLIP for few-shot classification," in *Proc. ECCV*. Springer, 2022, pp. 493–510.

[10] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan, "MaPLe: Multi-modal prompt learning," in *Proc. CVPR*, 2023, pp. 19113–19122.

[11] Sachit Menon and Carl Vondrick, "Visual classification via description from large language models," in *Proc. ICLR*, 2023.

[12] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie, "SuS-X: Training-free name-only transfer of vision-language models," in *Proc. ICCV*, 2023, pp. 2725–2736.

[13] Hai Zhang, Junzhe Xu, Shanlin Jiang, and Zhenan He, "Simple semantic-aided few-shot learning," in *Proc. CVPR*, 2024, pp. 28588–28597.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022, pp. 10684–10695.

[15] Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. NeurIPS*, 2018, vol. 31.

[16] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, "Meta-learning with differentiable convex optimization," in *Proc. CVPR*, 2019, pp. 10657–10665.

[17] Guneet S. Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto, "A baseline for few-shot image classification," in *Proc. ICLR*, 2020.

[18] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proc. CVPR*, 2020, pp. 8808–8817.

[19] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proc. CVPR*, 2020, pp. 12203–12213.

[20] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan, "Low-shot learning from imaginary data," in *Proc. CVPR*, 2018, pp. 7278–7286.

[21] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao, "Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners," in *Proc. CVPR*, 2023.

[22] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond, "Rethinking generalization in few-shot classification," in *Proc. NeurIPS*, 2022, vol. 35, pp. 3582–3595.

[23] Fusheng Hao, Fengxiang He, Jian Cheng, and Dacheng Tao, "Class-aware patch embedding adaptation for few-shot image classification," in *Proc. ICCV*, 2023, pp. 18905–18915.

[24] Fengyuan Yang, Ruiping Wang, and Xilin Chen, "Semantic guided latent parts embedding for few-shot learning," in *Proc. WACV*, 2023, pp. 5447–5457.

[25] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan, "Semantic prompt for few-shot image recognition," in *Proc. CVPR*, 2023, pp. 23581–23591.

[26] Delong Chen, Ji-Jia Zhang, Wei-Shi Zheng, and Ruixuan Wang, "Featwalk: Enhancing few-shot classification through local view leveraging," in *Proc. AAAI*, 2024.

[27] Stylianos Poulakakis-Daktylidis and Hadi Jamali-Rad, "Beclr: Batch enhanced contrastive few-shot learning," in *Proc. ICLR*, 2024.

[28] Ming Pan, Haoyan Xin, and Heng Tao Shen, "Semantic-based implicit feature transform for few-shot classification," *IJCV*, vol. 132, no. 11, pp. 5014–5029, 2024.

[29] Wenhao Li, Qiangchang Wang, Peng Zhao, and Yilong Yin, "KNN transformer with pyramid prompts for few-shot learning," in *Proc. ACM MM*, 2024, pp. 1082–1091.

[30] Yiman Hu, Yixiong Zou, Ruixuan Li, and Yuhua Li, "Generate universal adversarial perturbations for few-shot learning," in *Proc. NeurIPS*, 2024, vol. 37, pp. 102672–102695.

[31] Mushui Liu, Fangtai Wu, Bozheng Li, Ziqian Lu, Yunlong Yu, and Xi Li, "Envisioning class entity reasoning by large language models for few-shot learning," in *Proc. AAAI*, 2025.

[32] Javier Rodenas, Eduardo Aguilar, and Petia Radeva, "Slot attention-based feature filtering for few-shot learning," *arXiv preprint arXiv:2508.09699*, 2025.

[33] Ji Wu, Shipeng Wang, and Jian Sun, "AMMD: Attentive maximum mean discrepancy for few-shot image classification," *Pattern Recognit.*, vol. 155, pp. 110680, 2024.