DZGCoOp: Dual-Branch Zero-Shot Guided Context Optimization for Few-Shot Image Classification

Anonymous CVPR submission

Paper ID *****

Abstract

Prompt learning has become the prevailing paradigm for adapting vision-language models such as CLIP to few-shot image classification. However, existing multi-branch prompt methods fuse branch predictions with fixed equal weights at the logit level, which is redundant for samples that the frozen CLIP already classifies with high confidence yet unable to reconcile the large disagreements that arise on low-confidence samples. Moreover, parallel branches typically share identical objectives, data, and losses, leaving the division of labor among branches implicit, while prompt injection restricted to the encoder input layer fails to exploit the hierarchical representations of deep transformers. We propose DZGCoOp, a Dual-Branch Zero-Shot Guided Context Optimization framework. DZGCoOp first distills frozen-CLIP priors—LLM-enriched semantic anchors, zero-shot visual features, a zero-shot probability distribution, and an instance-level confidence—then learns an asymmetric dual-branch multimodal prompt network in which a strongly guided, hand-crafted-initialized branch preserves pre-trained knowledge while a weakly guided, randomly-initialized branch explores task-specific semantics. The two branches, coupled through shared deep prompt layers on both encoders, are fused per-instance with the zero-shot confidence, and zero-shot guidance losses align learned features and output distributions with the frozen priors. An epoch-weighted averaging strategy further consolidates parameter snapshots along the training trajectory. On 11 benchmarks, DZGCoOp attains 80.33% average harmonic mean in base-to-novel generalization (+1.78% over MaPLe), and it also achieves consistent gains in cross-domain (60.66%) and cross-dataset (65.99%) transfer. Our code is available at: https://github.com/RainBus/DZGCoOp.

1. Introduction

Vision-language models (VLMs) such as CLIP [12] and ALIGN [7] learn aligned image-text embedding spaces from web-scale pairs and exhibit remarkable zero-shot classification: an image is categorized by comparing its embedding with class-name embeddings produced from hand-crafted templates such as "a photo of a {class}". Adapting such models to downstream tasks by full fine-tuning is expensive and prone to catastrophic forgetting, which has made prompt learning [18] the dominant few-shot adaptation paradigm: only a small set of continuous context vectors is optimized while both encoders stay frozen.

Despite steady progress, two structural limitations persist in current prompt learners. (i) Sample-agnostic branch fusion. Multi-branch or ensemble-style methods—including our prior MSGCoOp [15], CoCoOp [17], and AAPL [10]—aggregate branch logits with equal weights. For samples that frozen CLIP already predicts with high confidence, the branches largely agree and equal-weight averaging wastes both computation and representational capacity; for samples with high zero-shot uncertainty, the branches disagree substantially, and a plain mean cannot integrate their complementary evidence. The fusion weight is moreover identical for every instance, ignoring how reliably CLIP "knows" each sample. (ii) Implicit division of labor among branches. Parallel branches are typically trained with the same objective, data, and losses. Regularizers such as the diversity loss of MSGCoOp only implicitly discourage branch similarity, so branches may still converge to near-identical semantics, and none of them is explicitly responsible for preserving pre-trained knowledge versus exploring task-specific ones. In addition, text-side prompt learning that inserts vectors only at the encoder input [16, 18] underuses the hierarchical representations of deep transformers, whereas VPT [8] and MaPLe [9] show that injecting prompts across layers of both encoders yields

finer-grained adaptation.

This paper proposes DZGCoOp (Dual-Branch Zero-Shot Guided Context Optimization), which turns the frozen CLIP itself into a per-sample routing and supervision signal. A zero-shot guidance module first distills four priors from the frozen encoders: LLM-enriched semantic anchors [2, 15], zero-shot visual features, a zero-shot probability distribution, and an instance-level confidence α given by the maximum zero-shot probability. On top of these priors, a dual-branch multimodal learning module builds an asymmetric architecture: a strong branch, initialized from a hand-crafted template and strongly regularized toward the semantic anchors, is conservative and preserves pre-trained knowledge; a weak branch, randomly initialized and weakly regularized, explores task-specific semantics. Both text branches share deep prompt layers coupled with deep visual prompts following MaPLe [9]. The final logits fuse the two branches per instance as $\alpha\,\mathbf{l}^{\text{str}} + (1-\alpha)\,\mathbf{l}^{\text{weak}}$, so reliable CLIP knowledge dominates on confident samples while exploratory representations take over on uncertain ones. Two guidance losses align learned features and output distributions with the zero-shot priors, and an epoch-weighted averaging (EWA) strategy consolidates parameter snapshots along the training trajectory with a Gaussian schedule.

Our contributions are summarized as follows:

- We propose a zero-shot guidance module that exploits the frozen CLIP encoders to produce LLM-enriched semantic anchors, zero-shot visual features, zero-shot probabilities, and instance-level confidence, providing stable priors and an adaptive fusion signal for prompt learning.
- We design an asymmetric dual-branch multimodal prompt network with shared deep prompt layers, in which strong and weak branches assume explicit complementary roles and are fused dynamically by the instance-level zero-shot confidence.
- We introduce zero-shot feature and probability guidance losses together with an epoch-weighted averaging strategy, aligning optimization with pre-trained knowledge and smoothing training stochasticity.
- Extensive experiments show that DZGCoOp achieves 80.33% average harmonic mean on 11 base-to-novel benchmarks (+1.78% over MaPLe), 60.66% average accuracy in cross-domain transfer, and 65.99% in cross-dataset transfer, with ablations validating each component.

2. Related Work

Prompt learning for vision-language models. CoOp [18] first replaces hand-crafted templates with learnable continuous context vectors for CLIP. CoCoOp [17] conditions the prompt on each input image to improve generalization to unseen classes. ProGrad [19] restricts prompt updates to directions consistent with the zero-shot CLIP gradient, and KgCoOp [16] regularizes learnable prompts toward the "a photo of a {class}" embeddings to mitigate forgetting. Beyond the text side, VPT [8] inserts learnable tokens into vision transformers, and MaPLe [9] jointly prompts both encoders across depth by mapping deep text prompts to deep visual prompts. AAPL [10] disentangles attribute-related tokens from augmentation noise. Our prior MSGCoOp [15] ensembles N parallel context vectors with LLM-generated semantic anchors and a diversity regularizer. These methods either fuse branch outputs with fixed equal weights or regularize branches only implicitly, and none exploits CLIP's zero-shot prediction itself to decide per sample how much each branch should contribute. DZGCoOp differs in using the frozen encoders to produce an instance-level confidence that routes between an explicitly conservative branch and an explicitly exploratory one, while inheriting the deep cross-modal prompting of MaPLe.

Knowledge preservation with zero-shot priors. Preserving pre-trained knowledge during few-shot adaptation is a central concern of prompt learning. KgCoOp [16] and ProGrad [19] anchor optimization to hand-crafted template embeddings; MSGCoOp [15] enriches such anchors with LLM-generated class descriptions [1, 2]. Distillation-style approaches additionally constrain output distributions. DZGCoOp extends this line along two axes: the zero-shot prior is used not only as a feature-space anchor but also as a distribution-level teacher via a KL guidance loss, and the zero-shot confidence actively participates in inference-time branch fusion rather than only in training-time regularization.

Weight averaging. Stochastic Weight Averaging (SWA) [6] averages checkpoints sampled in a late phase of training and finds flatter minima. Prompt learning, however, exhibits strong epoch-dependent behavior: models trained longer overfit the source classes and transfer worse [15]. Our epoch-weighted averaging (EWA) generalizes uniform averaging with a Gaussian weighting schedule whose center and width select the favorable phase per transfer regime, subsuming "early-checkpoint" selection without extra inference cost.

3. Method

We propose DZGCoOp, a prompt learning framework that exploits the frozen CLIP's zero-shot capability to guide an asymmetric dual-branch multimodal prompt network. Fig. 1 illustrates the overall pipeline: a zero-shot guidance stage builds priors from the frozen encoders, a dual-branch learning stage extracts deep-prompted visual and textual features and fuses them per instance, and a loss-and-optimization stage aligns the fused prediction with both labels and zero-shot priors. Throughout training, only prompt-related parameters are optimized.

3.1. Preliminaries

CLIP consists of a vision encoder $\phi(\cdot)$ and a text encoder $\theta(\cdot)$ trained contrastively on web-scale image-text pairs. For a task with N_c classes $\{\text{cls}_1, \ldots, \text{cls}_{N_c}\}$, zero-shot classification encodes the template "a photo of a [CLASS]" into text features $\mathbf{w}_i^{\text{txt}}$ and predicts via cosine similarity between $\mathbf{w}^{\text{img}} = \phi(I)$ and each $\mathbf{w}_i^{\text{txt}}$. Prompt learning replaces the fixed template with learnable context vectors while keeping ϕ and θ frozen. Following MaPLe [9], we insert prompts at both the input layer (shallow) and intermediate layers (deep) of both encoders.

3.2. Zero-Shot Guidance Module

The zero-shot guidance module distills four priors from the frozen CLIP, which are consumed later by the fusion rule and the guidance losses.

Class-level priors. Following MSGCoOp [15], a large language model [1] generates multiple visually grounded descriptions of each class covering appearance, color, shape, and texture. The descriptions of class cls_i form a set D_i; encoding them with the frozen text encoder and averaging yields the semantic anchor

$$\mathbf{w}_i^{\text{zst}} = \frac{1}{|D_i|} \sum_{d \in D_i} \theta(d), \tag{1}$$

a consensus representation of the class in CLIP's semantic space. Since anchors are instance-independent, they are precomputed once before training.

Instance-level priors. For an input image I, the frozen vision encoder yields the zero-shot visual feature $\mathbf{w}^{\text{zsv}} = \phi(I)$. Cosine similarities against all anchors form the zero-shot logits $l_i^{\text{zs}} = \text{sim}(\mathbf{w}^{\text{zsv}}, \mathbf{w}_i^{\text{zst}})$, which after softmax with temperature τ give the zero-shot probability distribution

$$p_i^{\text{zs}} = \frac{\exp(l_i^{\text{zs}}/\tau)}{\sum_{j=1}^{N_c} \exp(l_j^{\text{zs}}/\tau)}. \tag{2}$$

The instance-level confidence is the maximum zero-shot probability,

$$\alpha = \max_i p_i^{\text{zs}} \in [1/N_c,\, 1], \tag{3}$$

which measures how deterministically the frozen CLIP "knows" this sample: $\alpha \to 1$ indicates a confident zero-shot prediction, whereas $\alpha \to 1/N_c$ indicates a near-uniform distribution. On large-scale benchmarks such as ImageNet, the average confidence on correct classes lies around 0.6–0.8, making α a reliable per-sample signal. All computations in this module use frozen parameters and incur no gradient cost.

3.3. Dual-Branch Multimodal Learning Module

The learnable part of DZGCoOp is an asymmetric dual-branch network with deep cross-modal prompting.

Deep cross-modal prompting. On the vision side, shallow visual prompts $\mathbf{v}^{\text{vis}}$ are prepended to the patch tokens, and deep visual prompts $\text{prompt}_{\text{layer}_i}^{\text{vis}}$ are injected into intermediate transformer layers, yielding a prompted encoder $\tilde{\phi}(\cdot;\cdot)$ whose backbone stays frozen. On the text side, shallow contexts $\mathbf{v}^{\text{str}}$ and $\mathbf{v}^{\text{weak}}$ feed the strong and weak branches respectively, while the deep text prompts $\text{prompt}_{\text{layer}_i}^{\text{text}}$ are shared between the two branches. The visual pathway is single and shared: the visual feature $\mathbf{w}^{\text{vis}} = \tilde{\phi}(I; V^{\text{vis}})$ serves both branches.

Asymmetric branch design. The strong branch concatenates $\mathbf{v}^{\text{str}}$ (initialized from the hand-crafted template "a photo of a") with the class token and encodes it as $\mathbf{w}_i^{\text{str}} = \tilde{\theta}(\text{cls}_i; V^{\text{str}})$; the weak branch uses randomly initialized $\mathbf{v}^{\text{weak}}$ and yields $\mathbf{w}_i^{\text{weak}} = \tilde{\theta}(\text{cls}_i; V^{\text{weak}})$. The strong branch thus starts aligned with pre-trained knowledge and is trained under a strong anchoring constraint, so it preserves zero-shot generalization; the weak branch starts from a random point under a weak constraint, so it explores task-specific semantics that the template cannot express. Compared with implicit diversity regularization in multi-prompt ensembles [15], this initialization-plus-guidance asymmetry assigns the two branches explicit and complementary roles.

Confidence-guided dynamic fusion. Each branch scores the image against its class features,

$$l_i^{\text{str}} = \text{sim}(\mathbf{w}^{\text{vis}}, \mathbf{w}_i^{\text{str}}), \quad l_i^{\text{weak}} = \text{sim}(\mathbf{w}^{\text{vis}}, \mathbf{w}_i^{\text{weak}}), \tag{4}$$

and the final logit vector is an instance-level convex combination (Fig. 2):

$$\mathbf{l}^{\text{fin}} = \alpha \cdot \mathbf{l}^{\text{str}} + (1 - \alpha) \cdot \mathbf{l}^{\text{weak}}. \tag{5}$$

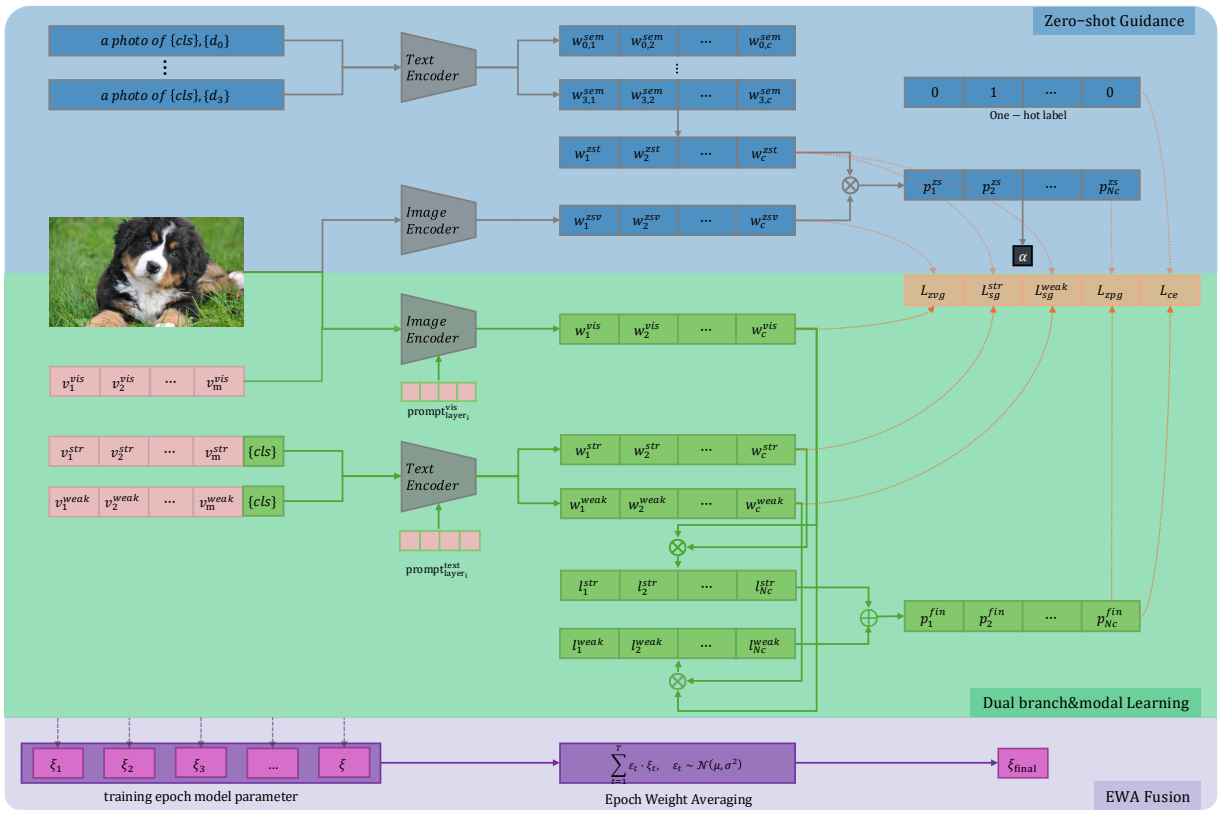


Figure 1. Architecture of DZGCoOp. Blue: the zero-shot guidance module distills LLM-enriched semantic anchors, zero-shot visual features, a zero-shot probability distribution, and an instance-level confidence α from the frozen CLIP encoders. Green: the dual-branch multimodal learning module couples a conservative strong branch and an exploratory weak branch through shared deep prompt layers on both encoders and fuses their logits with α. Purple & orange: guidance losses and the epoch-weighted averaging (EWA) strategy. Only prompt parameters are updated.

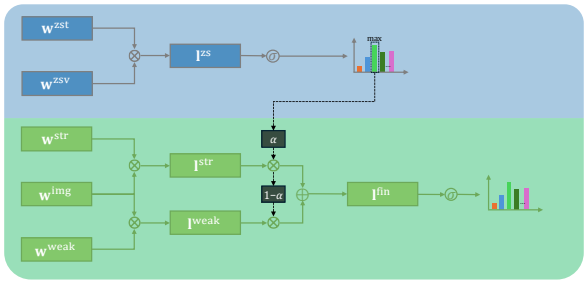


Figure 2. Zero-shot-confidence-guided dynamic fusion. The instance-level confidence α extracted from the frozen CLIP adaptively balances the conservative strong branch and the exploratory weak branch.

When CLIP is confident about a sample, the strong branch dominates and pre-training knowledge is preserved; when CLIP is uncertain, the weak branch contributes more of its task-specific evidence. Unlike equal-weight averaging, Eq. (5) requires no extra parameters and adapts the fusion weights to every sample. The fused logits are normalized by softmax into the final distribution $\mathbf{p}^{\text{fin}}$.

3.4. Guidance Losses and Optimization

Objective. The total loss combines classification and two zero-shot guidance terms:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{zfg}} + \mathcal{L}_{\text{zpg}}. \tag{6}$$

$\mathcal{L}_{\text{ce}}$ is the cross-entropy on the fused logits $\mathbf{l}^{\text{fin}}$. Because $\mathbf{l}^{\text{fin}}$ is the confidence-weighted combination of both branches, the classification gradient supervises them jointly.

Zero-shot feature guidance. To prevent the prompts from drifting away from CLIP's representation space, we align learned features with the zero-shot features via L_1 distances:

$$\mathcal{L}_{\text{zfg}} = \lambda_{\text{zvg}}\mathcal{L}_{\text{zvg}} + \lambda_{\text{sg}}^{\text{str}}\mathcal{L}_{\text{sg}}^{\text{str}} + \lambda_{\text{sg}}^{\text{weak}}\mathcal{L}_{\text{sg}}^{\text{weak}}, \tag{7}$$

$$\mathcal{L}_{\text{zvg}} = \frac{1}{B}\sum_{j=1}^{B}\|\mathbf{w}_j^{\text{vis}} - \mathbf{w}_j^{\text{zsv}}\|_1, \quad \mathcal{L}_{\text{sg}}^{\text{b}} = \frac{1}{N_c}\sum_{i=1}^{N_c}\|\mathbf{w}_i^{\text{b}} - \mathbf{w}_i^{\text{zst}}\|_1, \tag{8}$$

where B is the batch size and $\text{b} \in \{\text{str}, \text{weak}\}$. L_1 constrains feature positions dimension-wise rather than only directions. Mirroring the asymmetric design, the guidance weights satisfy $\lambda_{\text{sg}}^{\text{str}} > \lambda_{\text{sg}}^{\text{weak}}$ (with λ_{zvg} and $\lambda_{\text{sg}}^{\text{weak}}$ following the mature configuration of [15, 16]), so the strong branch is anchored tightly while the weak branch keeps freedom to explore.

Zero-shot probability guidance. At the distribution level, the zero-shot probabilities $\mathbf{p}^{\text{zs}}$ act as a soft teacher through a KL divergence [11]:

$$\mathcal{L}_{\text{zpg}} = \frac{1}{B \cdot N_c}\sum_{j=1}^{B}\sum_{i=1}^{N_c} p_{j,i}^{\text{zs}} \log \frac{p_{j,i}^{\text{zs}}}{p_{j,i}^{\text{fin}}}. \tag{9}$$

Beyond one-hot supervision, $\mathbf{p}^{\text{zs}}$ encodes CLIP's inter-class structure (semantically close classes receive similar probabilities), which regularizes few-shot training against overfitting and retains the ability to predict unseen classes.

Epoch-weighted averaging. Prompt learners conventionally use the last-epoch parameters, which are sensitive to stochastic fluctuation. As illustrated in Fig. 3, we snapshot the learnable parameters ξ_t at every epoch and fuse them with a Gaussian schedule after training:

$$\xi_{\text{final}} = \sum_{t=1}^{T} \epsilon_t \cdot \xi_t, \quad \epsilon_t \propto \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right), \tag{10}$$

where T is the number of epochs and (μ, σ) place and shape the weighting window. For base-to-novel generalization, later snapshots are of higher quality, so the window centers on late epochs (μ=15, σ=1); for cross-domain and cross-dataset transfer, early snapshots retain more generic pre-trained semantics, so the window shifts early and widens (μ=6, σ=10). Unlike SWA [6], which uniformly averages late-phase checkpoints and therefore hurts transfer when the model overfits the source dataset, EWA explicitly selects the favorable phase per task family. Inference uses ξ_{final} directly at no additional cost.

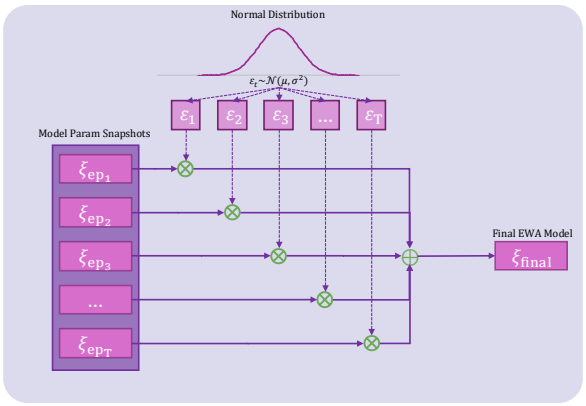


Figure 3. Epoch-weighted averaging (EWA). Snapshots along the training trajectory are fused with a Gaussian weight schedule whose center μ and width σ adapt to the transfer regime.

4. Experiments

4.1. Setup

Benchmarks and protocols. We evaluate three transfer protocols. (i) Base-to-novel generalization: following the standard splits, 11 datasets (ImageNet [3], Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, UCF101) are divided into base classes for few-shot training and novel classes for evaluation, reporting base accuracy (B), novel accuracy (N), and their harmonic mean (HM). (ii) Cross-domain transfer: the model trained on the 1000 classes of ImageNet is evaluated on ImageNetV2 [13], ImageNet-Sketch [14], ImageNet-A [5], and ImageNet-R [4]. (iii) Cross-dataset transfer: the ImageNet-trained model is directly evaluated on the other 10 datasets.

Implementation details. We use CLIP ViT-B/16 as the backbone. All models are trained in the 16-shot setting with batch size 4 using SGD (learning rate 0.0025) for 20 epochs with cosine annealing. Shallow visual and textual prompts have 4 tokens each. Deep prompt depths follow the recommended configuration of MaPLe [9]: 9 layers for base-to-novel and 3 layers for cross-domain and cross-dataset transfer. Guidance weights are λ_{zvg}=8 and $\lambda_{\text{sg}}^{\text{weak}}$=8, inherited from [15, 16], and $\lambda_{\text{sg}}^{\text{str}}$=24 (3× the weak weight) selected by grid search to realize the conservative-vs-exploratory asymmetry. EWA uses (μ, σ)=(15, 1) for base-to-novel and (6, 10) otherwise. All results are averaged over 3 random seeds.

4.2. Base-to-Novel Generalization

Table 1 reports the 16-shot results on the 11 benchmarks. DZGCoOp achieves the best averages of 84.52% (B), 76.54% (N), and 80.33% (HM), outperforming MaPLe by +2.24/+1.40/+1.78 and MSGCoOp by +3.12/+1.49/+2.23. It obtains the best HM on 8 of 11 datasets, with the largest gains on fine-grained and special-domain sets: FGVCAircraft (+3.52 HM over MaPLe), whose subtle inter-class differences benefit from the weak branch's exploratory features, Flowers102 (+3.65), StanfordCars (+3.09),

Table 1. Base-to-novel generalization on 11 benchmarks (16-shot). Bold marks the best result per column. Blue (resp. red) numbers show absolute gains (resp. drops) of DZGCoOp over MaPLe.

Method	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	B	N	HM	B	N	HM	B	N	HM	B	N	HM	B	N	HM	B	N	HM
CLIP	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12	63.37	74.89	68.65	72.08	77.80	74.83
CoOp [18]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47	78.12	60.40	68.13	97.60	59.67	74.06
CoCoOp [17]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43	70.49	73.59	72.01	94.87	71.75	81.71
AAPL [10]	80.27	72.17	76.01	76.53	70.57	73.43	97.87	95.10	96.46	95.63	97.40	96.51	70.33	73.50	71.88	95.10	70.63	81.06
KgCoOp [16]	80.73	73.36	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18	71.76	75.04	73.36	95.00	74.73	83.65
MSGCoOp [15]	81.40	75.05	78.10	76.46	70.45	73.31	98.08	95.81	96.91	95.62	97.89	96.70	71.22	74.56	72.85	96.20	75.91	84.84
MaPLe [9]	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58	72.94	74.00	73.47	95.92	72.46	82.56
DZGCoOp	84.52	76.54	80.33	77.80	70.74	74.11	98.30	95.05	96.65	95.67	97.61	96.63	78.35	74.85	76.56	97.69	77.14	86.21
	+2.24	+1.40	+1.78	+1.14	+0.20	+0.64	+0.56	+0.69	+0.63	+0.24	-0.15	+0.05	+5.41	+0.85	+3.09	+1.77	+4.68	+3.65

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	B	N	HM	B	N	HM	B	N	HM	B	N	HM	B	N	HM	B	N	HM
CLIP	90.10	91.22	90.66	27.19	36.29	31.09	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp [18]	88.33	82.26	85.19	40.44	22.30	28.75	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp [17]	90.70	91.29	90.99	33.41	23.71	27.74	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
AAPL [10]	90.70	91.60	91.15	34.07	24.17	28.28	79.65	76.90	78.25	73.90	53.43	62.02	87.00	66.30	75.25	82.20	74.27	78.03
KgCoOp [16]	90.50	91.70	91.09	36.21	33.55	34.83	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
MSGCoOp [15]	90.72	91.62	91.18	36.73	34.57	35.62	81.38	76.20	78.71	79.32	56.64	66.09	86.29	74.97	80.23	83.52	77.05	80.16
MaPLe [9]	90.71	92.05	91.38	37.44	35.61	36.50	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
DZGCoOp	90.54	91.52	91.03	43.48	37.07	40.02	82.62	79.01	80.77	83.72	60.02	69.92	94.05	78.35	85.49	87.49	80.51	83.85
	-0.17	-0.53	-0.35	+6.04	+1.46	+3.52	+1.80	+0.31	+1.02	+3.36	+0.84	+1.76	-0.02	+5.12	+3.14	+4.49	+1.85	+3.08

Table 2. Cross-domain transfer with 16-shot setting. The model is trained on ImageNet and evaluated on four shifted target domains.

	Source	Target				
	ImageNet	V2	Sketch	A	R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.90
AAPL	71.37	64.20	48.80	50.60	76.87	60.12
KgCoOp	71.20	64.10	48.97	50.69	76.70	60.11
MSGCoOp	71.59	64.54	49.24	50.90	76.94	60.41
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
DZGCoOp	71.66	64.77	49.43	51.37	77.08	60.66

Table 3. Cross-dataset transfer with 16-shot setting. The model is trained on ImageNet and directly evaluated on 10 downstream datasets.

Dataset	CLIP	CoOp	CoCoOp	AAPL	KgCoOp	MSGCoOp	MaPLe	DZGCoOp
ImageNet	66.72	71.51	71.02	71.37	70.66	71.59	70.72	71.66
Caltech101	92.94	93.70	94.43	94.17	93.92	93.29	93.53	94.69
OxfordPets	89.07	89.14	90.14	90.73	89.83	87.90	90.49	90.20
StanfordCars	65.29	64.51	65.32	65.10	65.41	65.10	65.57	65.51
Flowers102	71.30	68.71	71.88	71.67	70.01	69.87	72.23	71.00
Food101	86.11	85.30	86.06	86.00	86.36	85.60	86.20	86.36
Aircraft	24.87	18.47	22.94	23.03	22.51	19.14	24.74	22.89
SUN397	62.62	64.15	67.36	66.80	66.16	65.14	67.01	68.02
DTD	44.56	41.92	45.73	44.80	46.35	42.59	46.49	46.34
EuroSAT	47.69	46.39	45.37	41.83	46.04	45.59	48.06	44.82
UCF101	66.77	66.55	68.21	69.30	68.50	67.51	68.69	70.08
Average	65.12	63.88	65.74	65.34	65.51	64.17	66.30	65.99

EuroSAT (+3.14), and UCF101 (+3.08). Meanwhile, the strong branch keeps base accuracy high (best on 9 of 11 datasets), confirming that the asymmetric design improves novel-class generalization without sacrificing base-class fitness.

4.3. Cross-Domain Transfer

Table 2 shows robustness to distribution shift. DZGCoOp ranks first on all four target domains and on average (60.66%, +0.25 over MSGCoOp, +0.39 over MaPLe). The gain is most pronounced on the adversarial ImageNet-A (51.37%, +0.47), where the zero-shot guidance module supplies stable priors for samples far from the training distribution. We attribute the overall robustness to two designs: the frozen zero-shot branch remains a reliable reference under shift, and the EWA schedule configured for transfer (μ=6, σ=10) keeps early, more generic snapshots in the fused parameters, mitigating overfitting to the source domain.

4.4. Cross-Dataset Transfer

Table 3 evaluates direct transfer of the ImageNet-trained model. DZGCoOp reaches 65.99% on average, +1.82 over MSGCoOp and above CLIP, CoOp,

Table 4. Ablation of the zero-shot guidance module and losses.

Method	B	N	HM
IVLP	84.21	71.79	77.51
+ dual-branch + avg. fusion	84.33	72.10	77.74
+ ZG losses	84.38	76.38	80.18
+ dynamic fusion (full model)	84.52	76.54	80.33

Table 5. Ablation of the dual-branch design.

Method	B	N	HM
Single branch (weak only)	83.59	75.10	79.12
Single branch (strong only)	84.31	76.27	80.09
Symmetric dual (two strong) + avg.	84.65	75.50	79.81
Asymmetric dual (strong + weak) + avg.	84.38	76.38	80.18
Full (asym. dual + dynamic fusion)	84.52	76.54	80.33

CoCoOp, AAPL, and KgCoOp, though slightly below MaPLe (−0.31). It is the best on Caltech101 (94.69%), SUN397 (68.02%, +1.01 over MaPLe), and UCF101 (70.08%, +1.39). The confidence-guided fusion helps most on datasets whose categories are well covered by CLIP's pre-training (e.g., scenes and actions), while the residual gap to MaPLe on EuroSAT and Aircraft reflects the difficulty of satellite and fine-grained domains far from ImageNet semantics, where the source-trained weak branch provides limited additional evidence.

4.5. Ablation Studies

All ablations use the 16-shot base-to-novel setting averaged over the 11 datasets unless stated otherwise.

Zero-shot guidance. Table 4 starts from IVLP (independent visual-language prompting, i.e., MaPLe's prompting backbone without its coupling) [9]. Adding the asymmetric dual-branch with plain average fusion yields only +0.23 HM; the zero-shot guidance (ZG) losses are the key driver, lifting novel accuracy by +4.28 and HM by +2.44. Replacing average fusion with the confidence-guided dynamic fusion contributes a further +0.15 HM on top.

Asymmetric dual-branch. Table 5 dissects the branch structure. A single strong branch (80.09% HM) beats a single weak branch (79.12%), confirming the value of template initialization. Two strong branches with average fusion reach the highest base accuracy (84.65%) but a lower HM (79.81%) than the asymmetric pair (80.18%)—their predictions are highly correlated, so the ensemble degenerates into redundancy. The asymmetric strong-weak pairing improves novel accuracy by +0.88, and dynamic fusion adds another +0.15 HM, validating that the value of a second branch comes from complementary roles and instance-adaptive weighting rather than from ensembling per se.

Table 6. Ablation of the parameter fusion strategy.

Method	B	N	HM
Last-Epoch	84.18	76.14	79.96
SWA [6]	84.35	76.28	80.11
EWA (ours)	84.52	76.54	80.33

Parameter fusion strategy. Table 6 compares the final-parameter selection. Uniform late averaging (SWA) gains +0.15 HM over the last epoch; EWA's Gaussian schedule centered at μ=15 with σ=1 further reaches 80.33% (+0.37 over last-epoch, +0.22 over SWA), showing that explicitly weighting the stable late phase smooths single-epoch fluctuation. In cross-domain and cross-dataset regimes, the schedule shifts early (μ=6, σ=10), which is what preserves generic semantics in Tables 2 and 3.

5. Conclusion

We presented DZGCoOp, a zero-shot guided dual-branch prompt learning framework for few-shot image classification. DZGCoOp distills LLM-enriched semantic anchors, zero-shot visual features, probability distributions, and an instance-level confidence from the frozen CLIP; learns an asymmetric strong-weak dual-branch network coupled through shared deep prompt layers; fuses the branches per instance with the zero-shot confidence; and regularizes optimization with feature- and distribution-level guidance losses plus an epoch-weighted averaging strategy. Experiments on base-to-novel, cross-domain, and cross-dataset transfer demonstrate consistent improvements over strong prompt learning baselines, and ablations quantify the contribution of each component. We believe that exploiting a frozen VLM's own predictive confidence to route between conservative and exploratory adaptation paths is a general recipe for knowledge-preserving few-shot learning, and we leave its extension to denser vision-language tasks as future work.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62306320) and the Natural Science Foundation of Jiangsu Province (No. BK20231063).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Altman, Diogo Almeida, Max Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 2, 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, et al. Language models are few-shot learners. In Advances in Neural Information Processing Systems, pages 1877–1901, 2020. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. 5

[4] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8340–8349, 2021. 5

[5] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15262–15271, 2021. 5

[6] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, pages 876–885, 2018. 2, 5, 7

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, pages 4904–4916, 2021. 1

[8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In European Conference on Computer Vision, pages 709–727, 2022. 1, 2

[9] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19113–19122, 2023. 1, 2, 3, 5, 6, 7

[10] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1572–1582, 2024. 1, 2, 6

[11] Solomon Kullback and Richard A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 22(1):79–86, 1951. 5

[12] 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 Proceedings of the International Conference on Machine Learning, pages 8748–8763, 2021. 1

[13] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Proceedings of the International Conference on Machine Learning, pages 5389–5400, 2019. 5

[14] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In Advances in Neural Information Processing Systems, 2019. 5

[15] Zhaolong Wang, Tongfeng Sun, Mingzheng Du, Yachao Huang, Xinzheng Xu, and Zhongnian Li. MSGCoOp: Multiple semantic-guided context optimization for few-shot learning. arXiv preprint arXiv:XXXX.XXXXX, 2025. 1, 2, 3, 5, 6

[16] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6757–6767, 2023. 1, 2, 5, 6

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16816–16825, 2022. 1, 2, 6

[18] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130 (9):2337–2348, 2022. 1, 2, 6

[19] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15659–15669, 2023. 2