# TRUST WHAT YOU SEE OR WHAT YOU KNOW? EVIDENCE-COMPLEMENTARY PROTOTYPE PROMPTING FOR BIOMEDICAL VISION–LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Few-shot adaptation of biomedical vision–language models can draw on two kinds of evidence: text descriptions that know each class in general terms, and labeled images that show how the classes look in the target task. We observe that the two are reliable in opposite regimes: the text prior helps most when labeled images are scarce, while class prototypes built from the labeled images overtake it as the images accumulate. Most existing prompt learners, however, keep the influence of the text fixed no matter how many labels are available. A second problem appears when prototypes are simply added to a prompt trained with distillation: the distillation loss pulls their weight toward zero, because the text teacher has never seen the labeled images. We therefore propose Evidence-Complementary Prototype Prompting (ECPP), which lets each kind of evidence contribute where it is reliable. Concretely, (1) an *evidence handoff* gradually shifts trust from the text prior to a prototype classifier as more labeled images become available, and (2) *text-path distillation* keeps the teacher's guidance on the text branch, so that the prototype weight is learned from the labels alone; we show analytically why this placement matters. ECPP adds only one trainable scalar to a frozen BiomedCLIP prompt learner. On eleven biomedical datasets spanning nine imaging modalities, ECPP improves over the state-of-the-art BiomedCoOp at every support size (with gains of up to 5.34% on breast ultrasound) and achieves the best base-to-novel harmonic mean among the compared methods.

## 1 INTRODUCTION

Recognizing diseases from medical images is essential for many clinical applications, but expert annotation is tedious and costly, and rare classes may have only a handful of labeled examples. Few-shot adaptation of pretrained models is therefore increasingly attractive in medical image analysis. Vision–language models such as CLIP and BiomedCLIP map images and text into a shared feature space and classify an image by comparing it with the text feature of each class name (Radford et al., 2021; Zhang et al., 2023). In biomedical tasks, however, this zero-shot classifier is often not accurate enough, because classes may differ in subtle aspects of tissue texture, lesion boundary, or acquisition contrast that a class name does not convey.

Few-shot adaptation uses a small *support set* of $K$ labeled images per class to adapt a pretrained model to the target task. Prompt learning does this with both encoders frozen: CoOp trains only a few context tokens placed before the class name (Zhou et al., 2022b). The state-of-the-art BiomedCoOp further asks a large language model to describe each class (Koleilat et al., 2025), and uses the descriptions in two ways: as a fixed classifier that needs no training, which we call the *text prior*, and as a *teacher* that keeps the learned prompt close to the descriptions through knowledge distillation (see details in Section 3). The support images also give rise to a classifier that involves no text at all: averaging their features within each class gives a *class prototype*, and comparing a test image with the prototypes is a classifier in its own right (Snell et al., 2017; Zhang et al., 2022). The text prior and the prototype classifier thus offer two kinds of evidence: general knowledge about each class, and direct observations of the target task.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

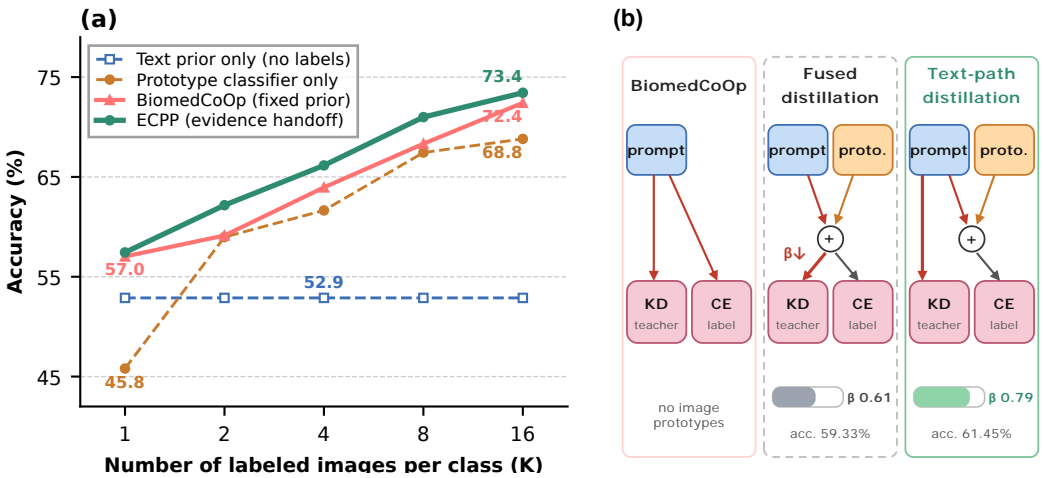



Figure 1: **Which evidence to trust, and where to distill.** (a) Few-shot accuracy on the BiomedCoOp benchmark, averaged over eleven datasets and three seeds. The text prior uses no labels; prototypes from two images use the refined centroid of Section 3.2; BiomedCoOp keeps the prior's influence fixed. **Prototypes overtake the text prior from two images per class on; ECPP leads BiomedCoOp at every $K$.** (b) Learned prototype weight $\beta$ (averaged over the benchmark) and few-shot accuracy (DermaMNIST) when the distillation loss is applied to the fused prediction or to the text path. **Fused distillation shrinks the prototype weight $\beta$; text-path distillation (ECPP) does not.**

Our starting point is the observation that these two sources of evidence are reliable in opposite regimes (Figure 1a). The text prior uses no labels, so its accuracy stays at 52.9%, averaged over eleven benchmark datasets, no matter how many labeled images are available. The prototype classifier is poor with a single labeled image per class (45.8%), because one image mixes the appearance of the class with that of the particular sample, but it overtakes the prior with two images and reaches 68.8% with sixteen. Ideally, then, the model should trust what it knows when labels are scarce, and what it sees as they accumulate. Existing prompt learners do not do this: the state-of-the-art BiomedCoOp keeps the influence of the text prior fixed regardless of $K$, and Figure 1a shows the cost. It improves over both single sources, yet it trails a model that shifts trust from the prior to the prototypes as labels accumulate (ECPP) at every support size, most clearly with two to eight images per class, where the two sources must be balanced most carefully.

A second problem arises in training (Figure 1b). A straightforward way to add prototypes to BiomedCoOp is to add their scores to the prompt's prediction and apply the distillation loss to this *fused* prediction. The teacher, however, knows only the text and has never seen the labeled images, so the loss penalizes the prototypes whenever they disagree with it, even when they fit the labels. In the simplest case the loss is minimized by removing the prototypes altogether (Section 3.4), and in practice the learned prototype weight drops from 0.79 to 0.61. The model thus loses exactly the evidence that the text cannot provide.

We therefore propose Evidence-Complementary Prototype Prompting (ECPP), which puts this principle into practice: the text prior and the prototypes each contribute where they are reliable. ECPP instantiates two designs in the BiomedCoOp framework (Figure 2), each targeting one of the two problems above. First, an *evidence handoff* gradually shifts trust from the text prior to the prototype classifier as more labeled images become available: a visual residual adds the prototype classifier with a learned weight, and prior attenuation reduces the weight of the text prior in the text-based prediction. Second, *text-path distillation* applies the distillation loss to the text-based prediction, before the prototypes are added (Figure 1b, right); the prompt is still kept close to the teacher, while the prototype weight is learned from the support labels alone. ECPP remains lightweight: it adds no new module, only one trainable scalar, and the handoff follows the size of the support set with no per-dataset tuning, so one set of constants serves all eleven datasets.

We evaluate ECPP on eleven biomedical datasets spanning nine imaging modalities. ECPP outperforms the state-of-the-art BiomedCoOp at every support size and on ten of the eleven datasets,

with gains of up to 5.34% on breast ultrasound, and achieves the best base-to-novel harmonic mean among the compared methods. Ablations show that both halves of the evidence handoff and text-path distillation contribute to these gains, and controls that give BiomedCoOp the same text prior at inference confirm that the improvement comes from better adaptation rather than from the prior itself. These gains cost little: ECPP introduces one additional parameter, with 2.8% more training time and 1.1% more inference time than BiomedCoOp.

## 2 Related Work

**Prompt learning for vision–language models.** CLIP classifies an image by its similarity to the text feature of each class (Radford et al., 2021). CoOp replaces the hand-written prompt by learned context tokens (Zhou et al., 2022b), CoCoOp makes the tokens depend on the image (Zhou et al., 2022a), and MaPLe learns prompts in both the text and the image encoder (Khattak et al., 2023a). Because a prompt trained on a few images can drift from what the pretrained model knows, KgCoOp constrains the learned embedding (Yao et al., 2023), ProGrad projects the gradient away from the zero-shot prediction (Zhu et al., 2023), and PromptSRC adds self-regularization and prompt ensembling (Khattak et al., 2023b). ECPP keeps such a regularizer, but asks where it should act once the prediction also contains prototypes that the pretrained text never saw.

**Biomedical vision–language models.** Biomedical pretraining aligns images with radiology reports or scientific text (Boecking et al., 2022; Wang et al., 2022; Zhang et al., 2023). BiomedCoOp adapts BiomedCLIP with prompt learning, using language-model descriptions of each class as a fixed classifier and as a distillation teacher (Koleilat et al., 2025); Section 3 gives the details. BiomedCCPL generates image-conditional prompts from multi-scale visual prototypes with a causal-disentanglement loss (Cui et al., 2026). Other recent directions include prompt policies that change with the input (Miao et al., 2026), prompt distributions on a hypersphere (Shao et al., 2026), and evidential tuning (Koleilat et al., 2026). ECPP instead keeps fixed prototypes computed once from the support images and studies how to combine them with a distillation-trained prompt.

**Prototype and adapter methods.** The mean feature of each class is a standard few-shot classifier (Snell et al., 2017; Chen et al., 2019). For CLIP, CLIP-Adapter learns a small feature transformation (Gao et al., 2024), Tip-Adapter builds a cache of support features and adds its votes to the zero-shot logits (Zhang et al., 2022), TaskRes learns a residual on top of the frozen text classifier (Yu et al., 2023), and LP++ is a linear probe initialized from the text features (Huang et al., 2024). Recent work aligns and mixes text and prototype representations without training (Goswami et al., 2026). ECPP combines these tools with a distillation-trained prompt: the weight of the text prior follows the number of labeled images, and the distillation loss is placed before the prototypes are added.

## 3 Evidence-Complementary Prototype Prompting

### 3.1 Background: prompt learning with a text prior and a teacher

Let $\mathcal{S} = \{(X_{ci}, Y_{ci})\}_{c=1,i=1}^{C,K}$ be the support set: $K$ labeled images $X_{ci}$ for each of $C$ classes, with labels $Y_{ci} = c$. The frozen image encoder maps an image to a unit-norm feature $f(X) \in \mathbb{R}^d$, and the frozen text encoder $h$ maps a text input to an unnormalized feature. Prompt learning trains four context tokens $\theta \in \mathbb{R}^{4\times 768}$, shared by all classes, and obtains the text feature of class $c$ as $t_{\theta,c} = \mathrm{norm}(h([\theta, c]))$, where $[\theta, c]$ is the context followed by the class name. BiomedCoOp also has $M = 50$ language-model descriptions $d_{mc}$ of each class, and averages their features into a second, fixed text feature $t_{0,c} = \mathrm{norm}(M^{-1}\sum_m h(d_{mc}))$. With the pretrained logit scale $s$, the two classifiers score a test image $X$ by

$$z_{\theta,c}(X) = sf(X)^\top t_{\theta,c}, \qquad z_{0,c}(X) = sf(X)^\top t_{0,c}. \tag{1}$$

We call $\mathbf{z}_\theta$ the *learned prompt* and $\mathbf{z}_0$ the *text prior*; the prior uses no labels and does not change during training. BiomedCoOp trains $\theta$ with three losses: cross-entropy on the support labels; a *semantic-consistency* term that keeps $t_{\theta,c}$ close to the description features; and a *knowledge-distillation* term that makes the class probabilities predicted by the prompt match those predicted by the descriptions,

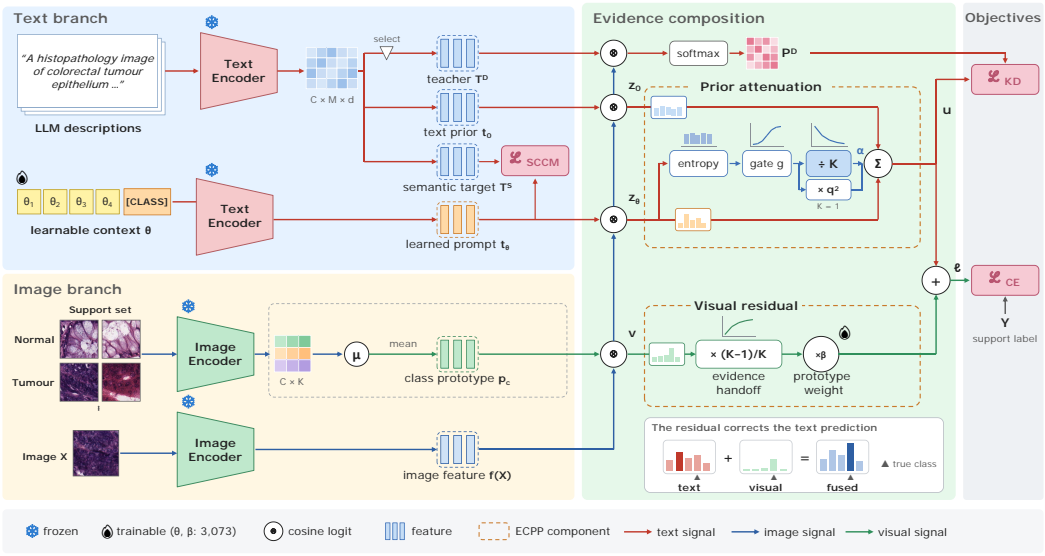



Figure 2: **ECPP overview.** The frozen text encoder turns the class descriptions into the text prior $t_0$, the semantic-consistency target $T^S$, and the distillation teacher $T^D$, and the learnable context into the prompt $t_\theta$; the frozen image encoder gives the feature $f(X)$ of an image $X$ and, once before training, the class prototypes $p_c$. The evidence handoff mixes the prior into the prompt with a weight that shrinks as $1/K$ (prior attenuation) and adds the prototype classifier with weight $\beta(K-1)/K$ (visual residual). The distillation loss acts on the mixed text logits $\mathbf{u}$, and the cross-entropy with the support label $Y$ on the fused logits $\boldsymbol{\ell}$. The inset is schematic.

which act as the *teacher*. The teacher uses a per-batch selection of the descriptions, the prior uses all of them, and both are frozen (Section 3.4).

ECPP keeps this training recipe and adds two designs (Figure 2). The *evidence handoff* consists of (i) a visual residual built from class prototypes (Section 3.2) and (ii) attenuation of the text prior (Section 3.3); as $K$ grows, the weight of the residual rises as $(K-1)/K$ and that of the prior falls as $1/K$. *Text-path distillation* applies the distillation loss before the residual is added (Section 3.4). Only $\theta$ and one scalar $\beta$ are trained.

### 3.2 Evidence handoff (i): visual residual from class prototypes

Intuitively, labeled images of a class show what the text cannot describe. Before training, we encode each support image once with the evaluation transform and average the features of each class into a unit-norm prototype $p_c = \mathrm{norm}(\sum_i f(X_{ci}))$. Comparing a test image with the prototypes gives a third classifier, $v_c(X) = s f(X)^\top p_c$, based on labeled images of the class rather than on text about it. We add it to the text-based prediction as a residual,

$$r_K = \frac{K-1}{K}, \qquad \text{visual residual} = \beta\, r_K\, \mathbf{v}(X), \tag{2}$$

where $\beta$ is a learned scalar initialized at 0.5. The factor $r_K$ switches the residual off at $K = 1$, where a prototype is a single image (Figure 1a), and approaches one as $K$ grows. At $K = 1$ the support image still trains the prompt through the classification loss. A prototype of two images is also noisy, so at $K = 2$ we refine it: the two images are weighted by their agreement with the text prior, and the result is interpolated toward $t_{0,c}$ with a weight chosen on the released validation subset before training. All other $K$ use the plain average. Appendix A gives the procedure.

### 3.3 Evidence handoff (ii): prior attenuation

Intuitively, the prior should matter most when labels are scarce. We make its weight depend on two things: how many labeled images there are, and how uncertain the learned prompt is about

the current image. Let $\pi(X) = \mathrm{softmax}(\mathbf{z}_\theta(X))$ and let $H(X) = -\sum_c \pi_c(X) \log \pi_c(X) / \log C$ be its normalized entropy, which is close to one when the prompt is uncertain. We set $g(X) = \sigma((H(X) - 0.5)/0.25)$ and mix the two text classifiers as

$$\alpha(X, K) = \begin{cases} q^2 g(X), & K = 1, \\ g(X)/K, & K \geq 2, \end{cases} \qquad \mathbf{u}(X) = (1 - \alpha)\mathbf{z}_\theta(X) + \alpha \mathbf{z}_0(X). \tag{3}$$

We call $\mathbf{u}$ the *mixed text logits*. For $K \geq 2$ the weight of the prior is at most $1/K$, so the prior can matter with two labeled images but contributes little with sixteen; the entropy term decides how much of this allowance a given image uses. At $K = 1$ there is no visual residual, and the allowance is instead scaled by $q = \mathrm{clip}((C^{-1} \sum_c p_c^\top t_{0,c} - 0.30)/0.15, 0, 1)$, which measures how well the single support image of each class agrees with the text prior. All constants are shared across datasets. The final prediction for a class with labeled images adds the visual residual:

$$\boldsymbol{\ell}(X) = \mathbf{u}(X) + \beta r_K \mathbf{v}(X). \tag{4}$$

Together, the two factors realize the evidence handoff: $r_K$ brings in the prototypes once there is more than one image per class, and $1/K$ lets the text prior fade as labels accumulate.

### 3.4 TEXT-PATH DISTILLATION

Intuitively, the teacher knows the class descriptions, while the support labels know the task. ECPP keeps BiomedCoOp's semantic-consistency and distillation losses but applies the distillation loss to the mixed text logits $\mathbf{u}$, not to the fused prediction $\boldsymbol{\ell}$. For a batch of $B$ images, training minimizes

$$\mathcal{L} = \frac{1}{B} \sum_i \mathrm{CE}(Y_i, \boldsymbol{\ell}_i) + \frac{\lambda_S}{Cd} \sum_c \|t_{\theta,c} - T_c^S\|_2^2 + \frac{\lambda_D}{BC} \sum_i D_{\mathrm{KL}}\big(P_i^D \parallel \mathrm{softmax}(\mathbf{u}_i)\big). \tag{5}$$

Here $T_c^S$ is the mean of the individually normalized description features of class $c$, and $P_i^D = \mathrm{softmax}(s f_{\mathrm{ZS}}(X_i)^\top T^D(B))$ is the teacher's prediction, computed from the descriptions selected for the batch and a frozen copy $f_{\mathrm{ZS}}$ of the image encoder with the same weights as $f$. The selection rule, the coefficients $\lambda_S$ and $\lambda_D$, and the selection threshold are inherited from BiomedCoOp without change (Appendix A). The KL temperature is one. Gradients reach $\theta$ through both $\mathbf{z}_\theta$ and the entropy term.

**Why the placement of the distillation loss matters.** Suppose the distillation loss were applied to the fused prediction $\boldsymbol{\ell}_i = \mathbf{u}_i + \beta r_K \mathbf{v}_i$ instead. Its derivative with respect to the residual weight would be

$$\frac{\partial \mathcal{L}_{\mathrm{KD,fused}}}{\partial \beta} = \frac{\lambda_D r_K}{BC} \sum_{i,c} \big[\mathrm{softmax}(\boldsymbol{\ell}_i)_c - P_{ic}^D\big] v_{ic}, \tag{6}$$

so the teacher would directly pull on $\beta$. With text-path distillation this derivative is zero, and $\beta$ is set by the classification loss alone; the two variants differ only in what the teacher constrains. Consider the case in which the prompt already agrees with the teacher, $P_i^D = \mathrm{softmax}(\mathbf{u}_i)$, with $\mathbf{u}_i$ and $\mathbf{v}_i$ held fixed and $r_K, \lambda_D > 0$. Then the fused distillation loss is a convex function of $\beta$ whose minimum is at $\beta = 0$, and the minimum is unique whenever some $\mathbf{v}_i$ varies across classes (Appendix B). That is, fused distillation prefers to remove the prototypes even when they help classification. Beyond this case, Section 4 measures the effect of the placement directly.

**Novel classes.** In base-to-novel evaluation (Section 4), base classes are predicted with Equation 4. Novel classes have no support images and therefore no prototypes; they are predicted with the fixed text-only mixture $\boldsymbol{\ell}^{\mathrm{novel}} = 0.1\,\mathbf{z}_\theta + 0.9\,\mathbf{z}_0$. The same rule is used for every model we train, so that comparisons on novel classes reflect training rather than the rule; Table S2 reports its sensitivity to the weight.

**Testing phase.** Given a test image $X$, one pass through the image encoder gives $f(X)$; the prototype scores are then $C$ dot products with the cached prototypes, and the label is $\arg\max_c \ell_c(X)$. No test-time labels, optimization, or extra encoder are needed. During training, each step encodes one augmented support batch and applies Equation 5.

# 4 EXPERIMENTS

## 4.1 BENCHMARK, PROTOCOLS, AND CONTROLS

We use the BiomedCoOp benchmark (Koleilat et al., 2025) with its released splits and support samples: BTMRI, BUSI, CHMNIST, COVID-19, CTKidney, DermaMNIST, KneeXray, Kvasir, LungColon, OCTMNIST, and RETINA (Appendix G). The eleven datasets cover nine imaging modalities, from MRI and CT to histopathology, endoscopy, and fundus photography. **Few-shot recognition** trains with $K \in \{1, 2, 4, 8, 16\}$ labeled images per class and three support seeds and tests on all classes. **Base-to-novel recognition** splits the classes of each dataset into a base half and a novel half, trains on 16 images per base class, and tests on both halves; the harmonic mean (HM) of base and novel accuracy is the summary metric. Ten datasets are used, since BUSI has too few classes to split. For our runs, HM is computed per run before averaging. For fair comparison, all models share the frozen BiomedCLIP backbone and the transforms, optimizer, schedule, and loss coefficients of BiomedCoOp. Training runs for 100 epochs (few-shot) or 50 epochs (base-to-novel), and the last model is evaluated without checkpoint selection (Appendix A).

To measure what each component contributes, we use *matched removals*: starting from the full ECPP model, we remove one component at a time and retrain with identical data, seeds, and hyperparameters, and compare the two models on the same test images. The three removals are (i) setting the visual residual to zero, (ii) dropping the $1/K$ factor in Equation 3 for $K \geq 2$, and (iii) moving the distillation loss from $\mathbf{u}$ to $\boldsymbol{\ell}$. Each removal is run on every dataset, $K$, and seed, giving 495 few-shot and 90 base-to-novel paired comparisons. Confidence intervals resample whole datasets 20,000 times, keeping their $K$ values and seeds together, and so reflect variation across tasks; Appendix D reports variation across seeds.

## 4.2 FEW-SHOT RECOGNITION

Table 1: **Few-shot recognition.** Accuracy (%) averaged over 11 datasets and three seeds for $K$ labeled images per class. Zero-shot references use no labels and do not depend on $K$; adapted methods are ordered by their mean over the five values of $K$. Bold and underlining mark the best and second-best adapted scores in each column.

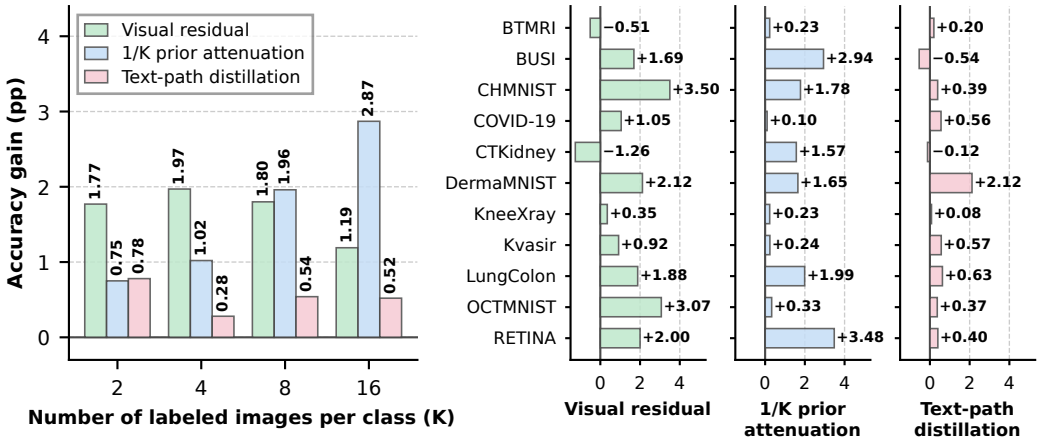
ZERO-SHOT REFERENCES (BiomedCLIP)

Single classifier: 42.05 Ensemble: 52.27 Selective ensemble: 53.72

| | Labeled examples per class | | | | | Overall |
|---|---|---|---|---|---|---|
| Method | 1 | 2 | 4 | 8 | 16 | Mean |
| CLIP-Adapter (Gao et al., 2024)[V] | 44.66 | 43.91 | 44.36 | 45.42 | 46.69 | 45.01 |
| CoCoOp (Zhou et al., 2022a)[P] | 48.49 | 51.28 | 54.69 | 61.08 | 65.09 | 56.13 |
| KgCoOp (Yao et al., 2023)[P] | 50.85 | 53.18 | 57.82 | 62.08 | 62.84 | 57.35 |
| Tip-Adapter (Zhang et al., 2022)[V] | 49.19 | 52.36 | 57.33 | 61.98 | 67.15 | 57.60 |
| LP++ (Huang et al., 2024)[L] | 47.24 | 53.18 | 59.02 | 63.69 | 68.35 | 58.30 |
| Standard LP[L] | 47.25 | 54.21 | 61.00 | 65.85 | 69.40 | 59.54 |
| CoOp (Zhou et al., 2022b)[P] | 50.16 | 54.18 | 59.75 | 65.84 | 69.62 | 59.91 |
| ProGrad (Zhu et al., 2023)[P] | 51.88 | 54.71 | 60.42 | 65.61 | 67.13 | 59.95 |
| Tip-Adapter-F (Zhang et al., 2022)[V] | 51.17 | 52.74 | 61.23 | 65.91 | 70.91 | 60.39 |
| BiomedCCPL (Cui et al., 2026)[P] | 51.83 | 57.88 | 60.08 | 66.13 | <u>72.76</u> | 61.74 |
| BiomedCoOp (Koleilat et al., 2025)[P] | <u>57.03</u> | <u>59.13</u> | <u>63.95</u> | <u>68.32</u> | 72.42 | <u>64.17</u> |
| **ECPP (ours)**[P] | **57.44** | **62.17** | **66.15** | **70.99** | **73.43** | **66.04** |

[V] CLIP-based visual adapters [L] Linear probing [P] Prompt learning.

As shown in Table 1, ECPP reaches 66.04% mean accuracy and consistently outperforms all competitors at every $K$. Its margin over BiomedCoOp's published numbers reaches 2.67 points with eight labeled images per class (70.99% versus 68.32%) and 2.20 points with four (66.15% versus 63.95%). With the evidence handoff, the model relies on task-specific appearance that the class descriptions do not convey, while the prior still guides the prompt when labels are scarce. The gain is not limited




(a) Every component adds accuracy at every $K > 1$; the $1/K$ prior matters more as $K$ grows.

(b) The $1/K$ prior helps on all 11 datasets; the residual and text-path distillation help on 9 of the 11.

Figure 3: **Every component helps, at different $K$ and on nearly every dataset.** Each value is the accuracy of the full ECPP model minus that of the model retrained without the named component. (a) Averaged over the eleven datasets and three seeds at each $K > 1$ (all three are inactive at $K = 1$). (b) Averaged over all $K$ and seeds.

to one modality: against our own BiomedCoOp runs, which use the same seeds and splits, ECPP improves the mean accuracy of ten of the eleven datasets, including ultrasound, histopathology, OCT, endoscopy, and fundus tasks. Appendix C lists every dataset.

### 4.3 Ablation study

Table 2: **Effect of each component.** Accuracy of the full ECPP model minus the accuracy of the model retrained without that component, in percentage points. The full model scores 66.04% few-shot accuracy (FS) and 78.00% base, 77.21% novel, and 76.18% HM in base-to-novel recognition. FS averages all $11 \times 5 \times 3$ runs; base-to-novel averages all $10 \times 3$ runs. All models use the same novel-class rule.

| Component | Δ FS | Δ Base | Δ Novel | Δ HM |
|---|---|---|---|---|
| Visual residual | +1.35 | +2.62 | -0.10 | +1.18 |
| Prior attenuation | +1.32 | +1.37 | +0.04 | +0.51 |
| Text-path distillation | +0.42 | +0.56 | -0.04 | +0.28 |

We ablate each design in turn: the first two questions below concern the two halves of the evidence handoff, and the third concerns text-path distillation. Table 2 summarizes the matched removals, and Figure 3 breaks them down by $K$ and by dataset. In few-shot recognition the three components are worth +1.35, +1.32, and +0.42 points, with 95% intervals $[0.53, 2.13]$, $[0.67, 2.00]$, and $[0.09, 0.83]$ across datasets; the corresponding macro-F1 gains are +2.17, +1.28, and +1.00 points.

**Does the visual residual add information beyond the text?** Removing the visual residual lowers accuracy at every $K \geq 2$, by 1.77, 1.97, 1.80, and 1.19 points at $K = 2, 4, 8, 16$, and on nine of eleven datasets (Figure 3). The largest effects are on CHMNIST, OCTMNIST, and DermaMNIST.

**Should the prior keep the same weight at every $K$?** Dropping the $1/K$ factor leaves the entropy term in place, so this removal tests the dependence on $K$ alone. It lowers accuracy on all eleven datasets, and the loss grows with $K$: 0.75 points at $K = 2$ and 2.87 points at $K = 16$. Without the factor, the prior keeps the same weight no matter how many labeled images there are, and overrides a prompt that has more evidence than it does.

**What if we distill the fused prediction?** Moving the distillation loss from $\mathbf{u}$ to $\boldsymbol{\ell}$ lowers accuracy at every $K \geq 2$ and on nine of eleven datasets, by 0.42 points overall and 1.00 point in macro-F1. The learned $\beta$ also shrinks: over the 132 few-shot runs with $K \geq 2$, its mean is 0.7930 under text-path distillation and 0.6087 under fused distillation, and it is smaller on nine of eleven datasets (0.7933 versus 0.5961 in the 30 base-to-novel runs, smaller on all ten datasets). This is expected from Equation 6: the teacher pulls directly on $\beta$.

### 4.4 BASE-TO-NOVEL RECOGNITION

Table 3: **Base-to-novel recognition.** Accuracy (%) averaged over ten datasets, with 16 labeled images per base class. "Same novel rule" predicts novel classes with ECPP's novel-class rule. Bold and underlining mark the largest and second-largest values.

| Method | Base | Novel | HM |
|---|---|---|---|
| BiomedCLIP | 47.84 | 65.42 | 53.81 |
| CoOp (Zhou et al., 2022b) | 73.85 | 64.75 | 67.23 |
| CoCoOp (Zhou et al., 2022a) | 72.26 | 67.03 | 67.22 |
| KgCoOp (Yao et al., 2023) | 68.36 | 64.08 | 64.61 |
| ProGrad (Zhu et al., 2023) | 71.67 | 66.93 | 67.43 |
| BiomedCoOp (Koleilat et al., 2025) | 76.26 | 73.92 | 75.07 |
| BiomedCoOp + same novel rule | 75.68 | **77.35** | <u>75.16</u> |
| BiomedCCPL (Cui et al., 2026) | **78.61** | 70.60 | 72.60 |
| **ECPP (ours)** | <u>78.00</u> | <u>77.21</u> | **76.18** |

Table 4: **Training versus the novel-class rule.** Base-to-novel accuracy (%), ten datasets, three seeds. Rows without the handoff are BiomedCoOp. Novel classes use the learned prompt $\mathbf{z}_\theta$ alone, or $0.1\mathbf{z}_\theta + 0.9\mathbf{z}_0$ when both are checked.

| **Training** | | | **Novel classes** | | | | |
|---|---|---|---|---|---|---|---|
| Handoff | KD | Text-path KD | $\mathbf{z}_\theta$ | $\mathbf{z}_0$ | Base | Novel | HM |
| ✗ | ✓ | ✓ | ✓ | ✗ | 75.68 | 71.31 | 71.72 |
| ✗ | ✓ | ✓ | ✓ | ✓ | 75.68 | 77.35 | 75.16 |
| ✓ | ✗ | – | ✓ | ✓ | 76.18 | 77.12 | 74.75 |
| ✓ | ✓ | ✗ | ✓ | ✓ | 77.44 | 77.25 | 75.90 |
| ✓ | ✓ | ✓ | ✓ | ✗ | 78.00 | 72.11 | 72.98 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 78.00 | 77.21 | 76.18 |

ECPP obtains 78.00% base, 77.21% novel, and 76.18% HM (Table 3). Among the designs (Table 2), the visual residual has the largest effect, +2.62 base and +1.18 HM, with a 95% HM interval of $[0.51, 2.15]$; prior attenuation adds +1.37 base and +0.51 HM; text-path distillation adds +0.56 base and +0.28 HM. The gains come from the base classes, which have prototypes; novel accuracy changes by less than 0.1 points under the shared novel-class rule.

Table 4 separates training from the novel-class rule. On novel classes the mixture with the prior lifts ECPP from 72.11% (learned prompt alone) to 77.21%, so we fix the rule to the mixture and vary only training (the full weight sweep is in Table S2). With the same mixture, BiomedCoOp reaches 75.16% HM, and ECPP (last row) is 1.02 HM points and 2.31 base points higher; the difference between the two methods is thus a difference in how well they adapt to the base classes. Distilling the fused prediction or dropping distillation lowers base accuracy by 0.56 and 1.82 points, while novel accuracy stays within 0.1 points.

### 4.5 COST

ECPP trains 3,073 parameters: four context tokens and the scalar $\beta$. Relative to BiomedCoOp (Table 5), it adds only 2.8% training time and 1.1% inference time, plus 2.30 seconds to build the prototype cache once; the prototype classifier adds only $C$ dot products with cached vectors per image. Appendices E and F report other rules for the prior's weight and alternative training objectives.

Table 5: **Training and inference cost.** RTX 3090, FP32, BTMRI, $K = 16$, $224 \times 224$ inputs. Train time is one forward/backward/update step with batch size 4; inference time is one batch of 100 test images including text encoding; data loading is excluded. Peak memory is measured during training.

| Method | Trainable | Train ms | Infer. ms | Images/s | Peak GiB | Cache s |
|---|---|---|---|---|---|---|
| BiomedCoOp | 3,072 | 58.72 | 237.21 | 421.6 | 1.762 | 0.00 |
| ECPP | 3,073 | 60.35 | 239.92 | 416.8 | 1.762 | 2.30 |

### 4.6 Results on each dataset

Figure 4 shows the few-shot results on each dataset for ECPP and for our own runs of BiomedCoOp and BiomedCCPL on the same seeds and splits. Averaged over $K$ and seeds, it is more accurate than BiomedCoOp on ten of the eleven datasets and more accurate than BiomedCCPL on ten of the eleven. Its largest gains over BiomedCoOp are on BUSI (+5.34 points), OCTMNIST (+3.97), and CHMNIST (+3.65); over BiomedCCPL they are on BUSI (+14.23) and DermaMNIST (+9.11). The advantage is largest when labels are scarce: for $K \leq 4$, ECPP leads the two methods by 2.01 and 5.33 points on average (1.68 and 2.76 for $K \geq 8$). Its standard error over seeds averages 1.77 points, the same as BiomedCoOp's (1.82).

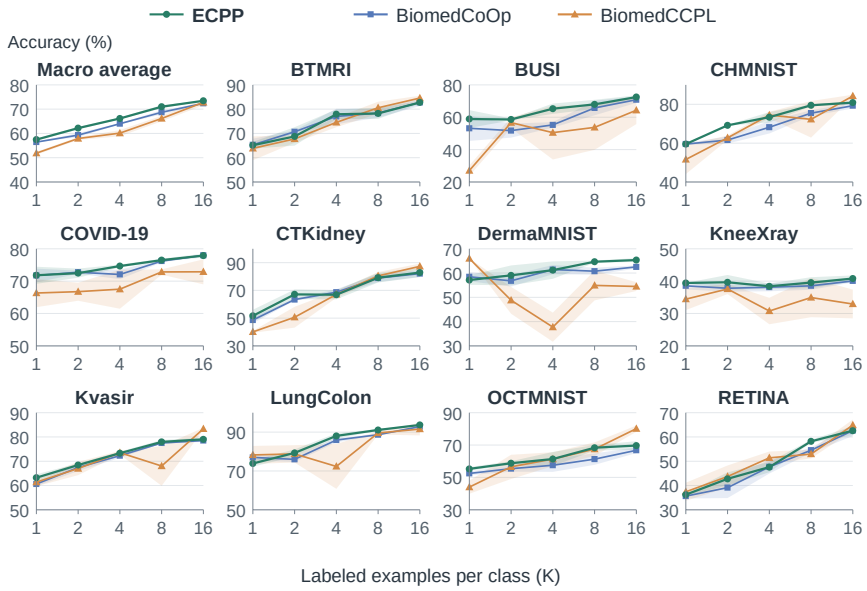



Figure 4: **Few-shot accuracy on each dataset.** Mean of three seeds with one-standard-error shading; all three methods are our own runs on the same seeds and splits. The macro panel averages datasets within each seed; vertical scales differ between panels.

## 5 Conclusion

We have presented ECPP, a method that lets text and visual evidence complement each other in few-shot biomedical adaptation. ECPP shifts trust from the text prior to a prototype classifier as labeled images accumulate, and keeps the distillation loss on the text path so that the teacher cannot suppress the prototypes; an analysis of the objective explains why the latter matters. Experiments on eleven datasets show consistent gains over the state-of-the-art BiomedCoOp (e.g., up to 5.34% on breast ultrasound), ablations confirm that each design contributes, and controls with a shared text prior show that the gain comes from adaptation. ECPP adds only one trainable scalar and keeps the cost of frozen-backbone prompt learning. In other words, the model should trust what it knows when labels are scarce, and what it sees as they accumulate. We hope ECPP serves as a plug-and-play recipe for adding visual evidence to distillation-trained prompt learners in medical image analysis and beyond. Code and trained prompts will be released upon acceptance.

## AI USE STATEMENT

Generative AI tools assisted with manuscript writing and revision. The authors are responsible for verifying the final research claims, code, results, and citations before submission.

## REFERENCES

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *European Conference on Computer Vision*, pp. 1–21, 2022.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Xueliang Cui, Juncai Zhang, Jiacheng Hou, Dan Lu, Hao Zhang, and Ruxin Wang. BiomedCCPL: Causal conditional prompt learning for biomedical vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40812–40821, 2026.

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. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Dipam Goswami, Simone Magistri, Gido M. van de Ven, Bartłomiej Twardowski, Andrew D. Bagdanov, Tinne Tuytelaars, and Joost van de Weijer. Cross-modal prototype alignment and mixing for training-free few-shot classification. *arXiv preprint arXiv:2603.24528*, 2026. URL https://arxiv.org/abs/2603.24528.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

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*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. BiomedCoOp: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14766–14776, 2025.

Taha Koleilat, Hassan Rivaz, and Yiming Xiao. Evi-Steer: Learning to steer biomedical vision-language models through efficient and generalizable evidential tuning. *arXiv preprint arXiv:2605.26292*, 2026. URL https://arxiv.org/abs/2605.26292.

Pingyi Miao, Xianlai Chen, Kai Sun, Yunbo Wang, Shuang Zhao, and Ying An. BioDPP: Dynamic prompt policy learning for biomedical vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8052–8060, 2026. doi: 10.1609/aaai.v40i10.37751.

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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Minye Shao, Sihan Guo, Xinrun Li, Xingyu Miao, Haoran Duan, and Yang Long. vMFCoOp: Towards equilibrium on a unified hyperspherical manifold for prompting biomedical VLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8851–8859, 2026. doi: 10.1609/aaai.v40i11.37839.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022. doi: 10.18653/v1/2022.emnlp-main.256.

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*, pp. 6757–6767, 2023.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023.

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 *European Conference on Computer Vision*, pp. 493–510, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. URL `https://arxiv.org/abs/2303.00915`.

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*, pp. 16816–16825, 2022a.

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, 2022b. doi: 10.1007/s11263-022-01653-1.

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*, pp. 15659–15669, 2023.