Few-Shot Biomedical Image Classification by Alignment of Independently Pretrained Encoders

Daniel Shalam [1] **Simon Korman** [1]

Abstract

Few-shot biomedical classification is dominated by methods built on jointly trained vision-language models such as BiomedCLIP, which rely on paired image-report corpora that are costly to assemble. In parallel, strong unimodal biomedical encoders, such as the self-supervised vision model RAD-DINO, are trained independently of text. We introduce a framework that aligns independently pretrained vision and text encoders using only the few-shot support set, combining a closed-form orthogonal Procrustes map with a lightweight flow-matching refinement. On the 11-dataset BiomedCoOp benchmark, replacing the jointly trained biomedical vision encoder with a general-purpose DINOv3 encoder matches or surpasses BiomedCoOp at higher shot counts. On VinDr-CXR multi-label chest X-ray classification, aligning RAD-DINO with BiomedCLIP text consistently outperforms linear probing of RAD-DINO across all shot counts. These results show that post-hoc alignment can reduce dependence on paired multimodal pretraining while enabling flexible combinations of independently trained unimodal models.

Biomedical data spans multiple modalities, including medical images and textual information from reports, annotations, and scientific literature. Vision-language models (VLMs) integrate image and text by learning a shared embedding space through large-scale contrastive pretraining. This paradigm, introduced by CLIP (Radford et al., 2021), has been extended to biomedicine by PubMedCLIP (Eslami et al., 2023) and BiomedCLIP (Zhang et al., 2023), which are trained on paired image-text data.

Given such a shared space, classification can be performed by comparing image features with text-derived class prototypes. This mechanism underlies many few-shot adaptation methods for frozen VLMs, including CLIP-Adapter (Gao et al., 2021), TIP-Adapter (Zhang et al., 2022), and CoOp (Zhou et al., 2022), which refine image features, text prompts, or support-set similarities while preserving the structure induced by multimodal pretraining. In biomedicine, BiomedCoOp (Koleilat et al., 2025) follows this line and achieves strong performance across diverse few-shot datasets.

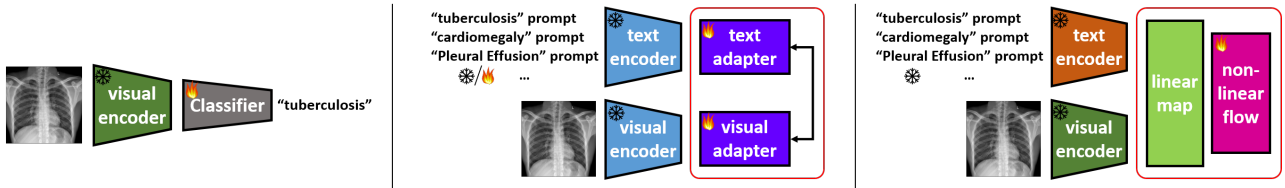
Paired image-text pretraining is more restrictive in biomedicine than in many natural-image settings (Zhang et al., 2023). Large, well-aligned image-report datasets are difficult to assemble due to privacy and access constraints. Moreover, reports are not complete descriptions of images: they emphasize findings relevant to the original interpretation and omit visual patterns that may matter for later tasks (P'erez-Garc'ia et al., 2024). Thus, contrastive image-text pretraining can align images to a sparse, task-dependent, and biased supervisory signal.

This has motivated renewed interest in *unimodal* encoders, trained independently within each modality rather than jointly across image-text pairs. RAD-DINO (P'erez-Garc'ia et al., 2024), for example, shows that a self-supervised vision encoder trained only on radiology images can match or surpass language-supervised biomedical encoders on classification and segmentation tasks. More broadly, strong unimodal encoders are now available in both vision and language, including DINOv3 (Siméoni et al., 2025) and LLM-derived text embeddings such as Qwen (Yang et al., 2025). These models avoid paired multimodal supervision, but are not aligned a priori and therefore do not provide the direct image-text comparison mechanism that makes VLMs useful for few-shot classification.

Fig. 1 summarizes the gap identified above and our proposed approach. A strong unimodal vision encoder can be frozen and used with a trainable linear classifier, yielding a *vision-only* model but losing the text-based prototype mechanism of VLMs. In the *multimodal* setting, vision and language encoders are *jointly* pre-trained on *paired* data to produce a shared embedding space, which can then be adapted downstream. Our approach instead bridges *inde-*

[1]Department of Computer Science, University of Haifa, Haifa, Israel. Correspondence to: Daniel Shalam <dani360@gmail.com>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.



(a) vision-only linear probing

(b) common adaptation of ***jointly*** pretrained (**multimodal**) encoders

(c) our adaptation of ***independently*** pretrained (**unimodal**) encoders

Figure 1. Three paradigms for few-shot image classification. (a) Vision-only approaches learn task-specific classifiers on top of image encoders using labeled data. (b) Multimodal approaches use jointly pretrained vision-language models (VLMs) that embed images and text into a shared space learned from large-scale paired image-text data, and adapt this space using few-shot supervision via prompts or feature adapters. (c) Our setting: independently pretrained unimodal encoders are not aligned. We construct a shared multimodal representation space post-hoc by learning cross-modal alignment from few-shot supervision, combining linear alignment with a flow-based refinement.

pendently pre-trained *unimodal* vision and text encoders by learning the missing alignment from few-shot supervision, enabling text-based classification without joint pretraining or large-scale paired image-text data.

This leaves the challenge of recovering cross-modal correspondence from few-shot supervision while keeping the encoders frozen. Prior work on modular encoder alignment bridges independently trained vision and language models with learned projections (Zhai et al., 2022; Li et al., 2023a; Zhang et al., 2025), but relies on paired corpora larger than available in few-shot adaptation. Closer to our setting, linear alignment methods, including orthogonal Procrustes mappings (Merullo et al., 2023; Ouali et al., 2023), recover correspondence with little or no training, but cannot model residual non-linear mismatch between modalities.

We combine a closed-form linear alignment with a lightweight flow-matching refinement. The linear stage maps text prototypes into image feature space while preserving their geometry, and the flow then learns a smooth transport between text and image representations. While flow matching (Lipman et al., 2023; Liu et al., 2023) has primarily been used for generation (Liu et al., 2025; Li et al., 2025), we use it here as a compact alignment prior for few-shot classification.

Our contributions are:

- We introduce a framework for constructing a shared representation space from independently pretrained vision and text encoders using only few-shot supervision. The approach combines a closed-form linear alignment with a lightweight flow-based refinement, enabling post-hoc cross-modal alignment without paired image-text data or encoder modification.
- On the BiomedCoOp (Koleilat et al., 2025) biomedical benchmark, we demonstrate that a general-purpose DINOv3 vision encoder can replace the jointly trained biomedical vision encoder, matching BiomedCoOp at 4-shot and clearly outperforming it at 16-shot despite weaker 1-shot performance.
- On VinDr-CXR (Nguyen et al., 2020), we post-hoc align a vision-only encoder (RAD-DINO) with a text encoder (BiomedCLIP-B/16) using only a few positive examples per class, and consistently outperforming linear probing of the same image encoder across all shot counts, demonstrating the value of cross-modal alignment in a realistic multi-label clinical setting.

1. Method

Our method is a modular adaptation framework for few-shot image classification with independently trained image and text encoders, summarized in Fig. 2 (Appendix B). It first aligns the two embedding spaces using a closed-form Orthogonal Procrustes (OP) map, and then refines this alignment with a lightweight flow-matching module trained in the aligned feature space. At inference time, image and text representations are transported toward intermediate points and classified by similarity-based matching. The training and inference procedures are summarized in Algorithms 1 and 2 in Appendix A.

1.1. Problem Setup

We consider a C-way K-shot classification task with a labeled support set $\Omega = \{(s_i, l_i)\}_{i=1}^{N}$, where $N = C \cdot K$, and $l_i \in \{1, \ldots, C\}$. At test time, the model is evaluated on unlabeled query samples from the same class set.

Let $f_{\text{img}} : \mathcal{S} \to \mathbb{R}^{D_{\text{img}}}$ and $f_{\text{text}} : \mathcal{L} \to \mathbb{R}^{D_{\text{text}}}$ be frozen image and text encoders. Each image s and label l is mapped to unit-normalized embeddings

$$\boldsymbol{x} = \frac{f_{\text{img}}(s)}{\|f_{\text{img}}(s)\|}, \qquad \boldsymbol{y} = \frac{f_{\text{text}}(\hat{l})}{\|f_{\text{text}}(\hat{l})\|}, \tag{1}$$

where $\hat{l}$ denotes a fixed template sentence containing l. This formulation allows combining independently pretrained encoders, enabling the use of domain-specific uni-modal models that may be trained on larger or more relevant datasets than jointly trained multi-modal models.

1.2. Orthogonal Procrustes (OP) Alignment

We align the text and image embedding spaces via a semi-orthogonal linear map $\mathbf{W} \in \mathbb{R}^{D_{\text{img}} \times D_{\text{text}}}$, estimated from paired latent features. Given normalized embeddings $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, we solve the Orthogonal Procrustes problem

$$\mathbf{W}^\star = \underset{\mathbf{W}}{\operatorname{argmin}} \left\|\mathbf{T}\mathbf{W}^\top - \mathbf{X}\right\|_F^2 \quad \text{s.t.} \quad \mathbf{W}\mathbf{W}^\top = \mathbf{I}, \tag{2}$$

where $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ and $\mathbf{T} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^\top$.

Let $\mathbf{M} = \mathbf{X}^\top\mathbf{T}$ be the cross-covariance matrix, and denote its singular value decomposition by $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. The optimal mapping is given in closed form by $\mathbf{W}^\star = \mathbf{V}\mathbf{U}^\top$.

We map text embeddings $\boldsymbol{y} \leftarrow \mathbf{W}^\star\boldsymbol{y}$ into the image latent space, while keeping $\boldsymbol{x}$ fixed. This provides a geometry-preserving coarse alignment between modalities, serving as a stable initialization for subsequent refinement.

1.3. Flow-Matching Refinement

We refine the OP-aligned representations by learning a continuous transport between image embeddings and their corresponding class prototypes in the latent space. Given a pair $(\boldsymbol{x}, \boldsymbol{y})$ after OP alignment, we consider a time-indexed path $\gamma(t)$ with $t \in [0, 1]$ connecting $\boldsymbol{x}$ to $\boldsymbol{y}$, and denote by $\boldsymbol{x}_t = \gamma(t)$ the intermediate point along this path.

We use a spherical geodesic path $\gamma(t) = \operatorname{slerp}(\boldsymbol{x}, \boldsymbol{y}; t)$ between each support embedding and its OP-aligned class prototype. This path respects the unit-sphere geometry of the normalized features (Chen & Lipman, 2023).

We parameterize a time-dependent velocity field $v_\theta(t, \boldsymbol{x}_t)$ and train it to match the target velocity along the path. Sampling $t \sim \mathcal{U}[0, 1]$, the flow-matching objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y),\,t}\left[\|v_\theta(t, \boldsymbol{x}_t) - \dot{\gamma}(t; \boldsymbol{x}, \boldsymbol{y})\|_2^2\right]. \tag{3}$$

This yields a non-linear refinement of the OP alignment in the latent space. We train bidirectional flows by swapping the roles of $(\boldsymbol{x}, \boldsymbol{y})$, and use both directions during inference.

1.4. Inference

At test time, given a query image s, we compute its embedding $\boldsymbol{x}$ and the OP-aligned class prototypes $\boldsymbol{y}_c \leftarrow \mathbf{W}^\star\boldsymbol{y}_c$ for each class c.

We transport both $\boldsymbol{x}$ and $\boldsymbol{y}_c$ in the latent space using the learned flow to intermediate times τ and $1 - \tau$, respectively, yielding $\boldsymbol{z}^{(\tau)}$ and $\boldsymbol{z}_c^{(1-\tau)}$. This symmetric transport yields a balanced comparison between modalities, avoiding bias toward either endpoint. Classification is based on cosine similarity:

$$s_c = \left\langle \boldsymbol{z}^{(\tau)}, \boldsymbol{z}_c^{(1-\tau)} \right\rangle. \tag{4}$$

Table 1. **Few-shot classification on BiomedCoOp benchmark.** Accuracy (%), averaged over 11 datasets. Vision (Vis.) and text (Txt.) encoders are indicated for each method. BMC denotes BiomedCLIP-B/16, and DIN denotes DINOv3-B. **Bold** and *italics* denote best and second-best results per shot setting, respectively.

category	method	Vis. / Txt.	K=1	K=4	K=16
vis. only	Linear probe	BMC / –	47.25	61.00	69.40
joint pair	CLIP-Adapter	BMC / BMC	44.66	44.36	46.69
	Tip-Adapter-F	BMC / BMC	51.17	61.23	70.91
	CoOp	BMC / BMC	50.16	59.75	69.62
	ProGrad	BMC / BMC	51.88	60.42	67.13
	BiomedCoOp	BMC / BMC	**57.03**	*63.95*	72.42
indep. pair	**ours**	DIN / BMC	50.94	63.17	*73.71*
combined	**ours**	DIN / BMC	*52.01*	**64.14**	**74.41**

Following prior work (Gao et al., 2021; Lin et al., 2023), we combine the flow-based score s_c^{flow} with the base similarity $s_c^{\text{base}} = \langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$ via a convex combination

$$s_c = (1 - \alpha)\, s_c^{\text{flow}} + \alpha\, s_c^{\text{base}}, \tag{5}$$

where $\alpha \in [0, 1]$ controls the trade-off between the terms.

The score-mixing mechanism also allows us to combine complementary signals from independently trained and jointly trained encoder pairs. An independently trained pair can benefit from a strong vision encoder, such as DINOv3, while a jointly trained pair, such as BiomedCLIP, provides a robust zero-shot image-text prior. We therefore define a variant, categorized *combined*, by replacing s_c^{base} in Eq. (5) with the zero-shot similarity from a jointly trained pair:

$$s_c^{\text{base}} = \left\langle \boldsymbol{x}^{\text{joint}}, \boldsymbol{y}_c^{\text{joint}} \right\rangle, \tag{6}$$

where $\boldsymbol{x}^{\text{joint}}$ and $\boldsymbol{y}_c^{\text{joint}}$ are the BiomedCLIP-encoded image and class-prompt features. Together, these components define a simple yet expressive alignment framework for few-shot classification in the latent space.

2. Experiments

Baseline methods include CLIP-Adapter (Gao et al., 2021), CoOp (Zhou et al., 2022), TIP-Adapter/(F) (Zhang et al., 2022), ProGrad (Zhu et al., 2022) and BiomedCoOp (Koleilat et al., 2025).

2.1. Biomedical Few-Shot Classification

We evaluate our method on the BiomedCoOp benchmark (Koleilat et al., 2025), which comprises 11 biomedical image datasets spanning modalities such as X-ray, MRI, histopathology, and endoscopy. The benchmark follows a few-shot protocol, and we report representative shot counts of 1, 4 and 16.

BiomedCoOp builds on BiomedCLIP (Zhang et al., 2023), a jointly trained biomedical vision-language model pretrained on approximately 15M paired biomedical image-text samples. It adapts this backbone using a multi-stage pipeline with prompt learning, LLM-based prompt generation and ensembling, and knowledge distillation. In contrast, our work decouples the choice of encoders: we use the independently trained DINOv3-B vision encoder and align it with the BiomedCLIP-B/16 text encoder. This keeps the model scale equal to the BiomedCLIP-based baselines, while replacing the jointly trained biomedical vision encoder with a strong general-purpose visual representation.

Table 1 reports average accuracy over the 11 datasets. At 1-shot, BiomedCoOp remains stronger, suggesting that the biomedical jointly trained pair and prompt-based adaptation are especially useful in the low-shot regime. As more support examples become available, the advantage of using DINOv3-B becomes clear. In the independent-pair setting, our method is competitive with BiomedCoOp at 4-shot and improves over it at 16-shot. The combined variant improves performance, reaching 64.14 at 4-shot and 74.41 at 16-shot, surpassing BiomedCoOp at both shot counts.

Even though DINOv3-B is not specialized to biomedical images and is not jointly trained with text, it provides a strong visual backbone that can be aligned effectively from few-shot supervision. Per-dataset results (Tab. 6 in Appendix E) show that the gains are not concentrated in a single dataset. Notably, we obtain these results without prompt learning, LLM-based ensembling, or additional biomedical image-text pretraining.

2.2. Multi-Label Classification on Chest X-Rays

Clinical chest X-rays present a challenging classification setting. Images are grayscale with subtle and spatially diffuse cues, and in VinDr-CXR (Nguyen et al., 2020) supervision is *multi-label*, with imbalanced and potentially noisy labels extracted from radiology reports. We therefore evaluate using macro-AUPRC, a threshold-free ranking metric.

Following the protocol of RAD-DINO (P'erez-Garc'ia et al., 2024), we use their curated subset of VinDr-CXR annotated for 7 pathologies. Since the setting is multi-label, we define K as the minimum number of positive training examples per class, and drop all "no finding" images so that every training sample contributes at least one label. Following the protocol of RAD-DINO (P'erez-Garc'ia et al., 2024), we use their curated VinDr-CXR subset annotated for 7 pathologies. We construct a few-shot multi-label classification task by randomly sampling training subsets with at least K positive examples for each pathology, treating K as a per-class minimum because images may contain multiple labels. We discard "no finding" images, i.e., images that contain none of the 7 target pathologies, so that every sampled training image contributes at least one positive label. We use RAD-DINO as the image backbone. It is a vision-only self-supervised encoder trained on radiology images and represents the current state of the art on VinDr-CXR, outperforming even encoders trained with paired text supervision. For the text side we use the BiomedCLIP-B/16 text encoder (Zhang et al., 2023), and align the two encoders via our post-hoc alignment procedure. To handle multi-label supervision, each training image's label set is converted into a single target by averaging the normalized label prompts before OP alignment.

Table 2. **Few-shot multi-label classification on VinDr-CXR** Macro-AUPRC (%). Our method pairs RAD-DINO with BiomedCLIP-B/16 text by post-hoc alignment.

method	K=1	K=2	K=4	K=8	K=16
RAD-DINO probe	30.73	30.82	33.42	36.62	41.82
ours	**30.96**	**33.38**	**40.49**	**44.45**	**49.22**
Δ	+0.23	+2.56	+7.07	+7.83	+7.40

Table 2 compares our method against a linear probe of RAD-DINO features on the same K-shot support set. Our method matches the linear probe at K=1 and then grows a rapidly widening gap as K increases, reaching $+7.4$ pp at K=16. The alignment is especially effective in the low-shot regime where the linear probe is unstable, indicating that the text-conditioned flow provides an effective inductive bias under label noise and class imbalance. These results show that a uni-modal general-purpose vision encoder can be turned into a competitive multi-label classifier via post-hoc alignment to an independent text model, without any multimodal pretraining of the image encoder itself.

3. Conclusion

We presented a simple framework for few-shot biomedical image classification that aligns independently pretrained vision and text encoders using only the support set, without paired multimodal pretraining. On the BiomedCoOp benchmark, replacing the biomedical-specialized vision encoder with a general-purpose DINOv3 encoder matches or surpasses biomedical-specialized baselines at higher shot counts. On VinDr-CXR, aligning RAD-DINO with BiomedCLIP text consistently outperforms linear probing of the same image encoder. These results indicate that post-hoc alignment is a practical alternative to paired multimodal pretraining, making it easy to combine the strongest unimodal models as they emerge.

Impact Statement This work has no societal consequences that we believe require specific highlighting.

References

Chen, R. T. Q. and Lipman, Y. Flow matching on general geometries. In *International Conference on Learning Representations (ICLR)*, 2023.

Eslami, S., Meinel, C., and De Melo, G. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1151–1163, 2023.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. J. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 2021.

Koleilat, T., Asgariandehkordi, H., Rivaz, H., and Xiao, Y. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013.

Li, J., Li, D., Savarese, S., and Hoi, S. C. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023a.

Li, S., Kallidromitis, K., Gokul, A., Liao, Z., Kato, Y., Kozuka, K., and Grover, A. Omniflow: Any-to-any generation with multi-modal rectified flows. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Li, T., Katabi, D., and He, K. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2023b.

Lin, Z., Yu, S., Kuang, Z., Pathak, D., and Ramanan, D. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Liu, Q., Yin, X., Yuille, A., Brown, A., and Singh, M. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. In *International Conference on Learning Representations (ICLR)*, 2023.

Nguyen, H. Q., Lam, K., Le, L. T., et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9, 2020.

Ouali, Y., Bulat, A., Martinez, B., and Tzimiropoulos, G. Black box few-shot adaptation for vision-language models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

P'erez-Garc'ia, F., Sharma, H., Bond-Taylor, S., Bouzid, K., et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7:119 – 130, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Siméoni, O., Vo, H. V., Seitzer, M., et al. DINOv3, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision (ECCV)*, 2022.

Zhang, S., Dong, W., Xu, H., Yang, Y., He, X., and Yan, S. Freeze-align: Harnessing frozen unimodal encoders for flexible multimodal alignment. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

Zhang, Y., Xu, J., He, Z., et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15613–15623, 2022.

A. Algorithms

Algorithm 1 Training

Require: support set $\Omega = (\{\boldsymbol{s}_i\}_{i=1}^N, \{\boldsymbol{l}_i\}_{i=1}^N)$
Require: frozen encoders $f_{\text{img}}, f_{\text{text}}$
Require: initialized flow model v_θ
compute image and text embeddings $\{\boldsymbol{x}_i\}_{i=1}^N$ and $\{\boldsymbol{y}_i\}_{i=1}^N$ (Eq. 1)
estimate (globally or locally) an OP linear map $W^\star$ (Sec. 1.2)
align text to image using OP: $\boldsymbol{y}_i \leftarrow W^\star \boldsymbol{y}_i$ for $i = 1, \ldots, N$
for minibatches $(\boldsymbol{x}, \boldsymbol{y})$ **do**
  sample a batch of intermediate times $t \sim \mathcal{U}[0, 1]$
  obtain intermediate path points $\boldsymbol{x}_t = \gamma(t; \boldsymbol{x}, \boldsymbol{y})$
  compute target velocities $\boldsymbol{u}(t; \boldsymbol{x}, \boldsymbol{y})$ (linear or geodesic)
  update θ by minimizing the loss $\mathcal{L}_{\rightarrow}(\theta)$ (Eq. 3)
end for

Algorithm 2 Few-shot inference

Require: query s; class labels $\{l_c\}_{c \in C}$; encoders $f_{\text{img}}, f_{\text{text}}$; flow model v_θ; map $W^\star$; params τ, α;
compute embeddings $\boldsymbol{x}$ and $\{\boldsymbol{y}_c\}_{c \in C}$ (Eq. 1)
align text prototypes to image using OP: $\boldsymbol{y}_c \leftarrow W^\star \boldsymbol{y}_c$ for every $c \in C$
integrate ODE to obtain $\boldsymbol{z}^{(\tau)}$; symmetrically compute $\boldsymbol{z}_c^{(1-\tau)}$;
compute class scores s_c (Eq. 5)
predict class $c = \arg\max_{c \in C} s_c$

B. Our method's workflow

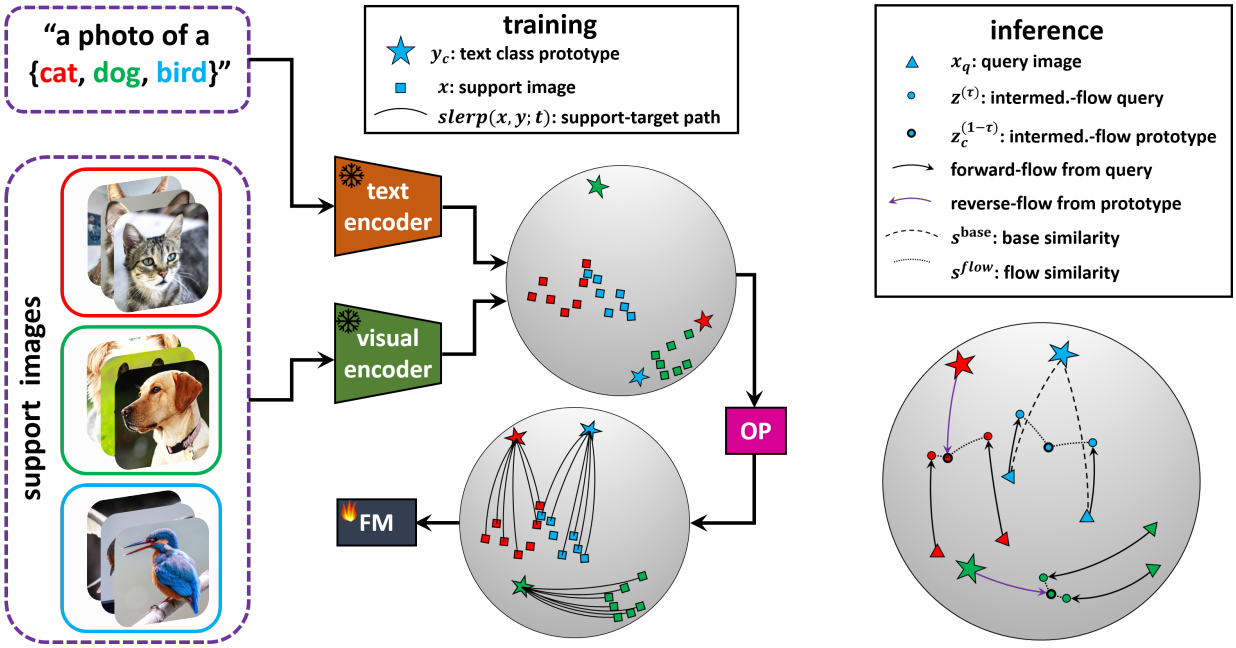


Figure 2. **Our method's workflow.** Support images and class labels are passed through frozen image and text encoders, producing unit-normalized features that are not aligned a priori. Orthogonal Procrustes (OP) provides a closed-form linear alignment of the text features to the image-feature space, giving a stable initialization for the shared latent space. A lightweight flow-matching (FM) module is then trained on paths between class prototypes and support-image features. At inference time, query-image features and class prototypes are integrated by the learned forward and reverse flows toward intermediate points, whose similarities are used for class prediction, together with the base OP-aligned similarity.

C. Implementation Details

We follow the evaluation protocols of the respective benchmarks. For the BiomedCoOp benchmark, we use the standard few-shot setup and report mean performance over 3 random splits. For VinDr-CXR, we construct few-shot training subsets with $K \in \{1, 2, 4, 8, 16\}$ minimum positive examples per class and evaluate using macro-AUPRC.

Our method first aligns frozen image and text encoders using an orthogonal Procrustes (OP) map, fit in closed form on the K-shot support set and kept fixed thereafter. We then learn a flow in the aligned latent space, parameterized by a velocity field $v_\theta(t, z)$ following Li et al. (2023b): a 2-layer residual MLP with hidden width 1536, `SiLU` activations, and per-layer time conditioning, with input and output dimensions matching the encoder dimensionality. Only the flow parameters are trained. Optimization uses AdamW with learning rate 5×10^{-5}, weight decay 10^{-3}, and a cosine learning rate schedule, for 200 epochs on all datasets except ImageNet, where we use 50 epochs. At inference, we transport query embeddings and class prototypes using a midpoint ODE solver with 2 steps.

D. Ablations

D.1. Necessity of the Linear Orthogonal Procrustes (OP) Alignment

This ablation highlights the importance of the OP module in our independently trained encoders setting. We ran our method on ImageNet, using a strong uni-modal pair (CLIP-L/14 text and DINOv2-B image) with matching feature dimensions (otherwise OP or some dimensionality reduction are anyway strictly required). The results show that our method can be entirely flow-based and work without the OP initial alignment, but this comes at a significant average cost of 2.41 in accuracy, with an expected widening of the gap towards the low-shot regime in which the flow model does not have enough data to learn the cross-modal alignment from scratch. This confirms that even when dimensions match, OP provides a critical coarse geometric alignment that stabilizes the flow-matching prior when supervision is scarce.

Table 3. **Impact of OP initialization** (with CLIP-L/14 + DINOv2-B).

Method	1-shot	2-shot	4-shot	8-shot	16-shot	Avg.
ours (w/o OP)	56.63	64.60	70.07	73.47	75.55	68.06
ours (w/ OP)	**60.73**	**67.47**	**72.17**	**75.00**	**77.00**	**70.47**
gain	+4.10	+2.87	+2.10	+1.53	+1.45	+2.41

D.2. Component Ablation

Table 4 ablates the main components of our method on ImageNet across shot counts, using DINOv3-B vision and CLIP-B/16 text. OP alone provides a useful but limited cross-modal bridge, staying below the CLIP zero-shot baseline at every shot count. Adding the flow refinement closes this gap and overtakes zero-shot CLIP, improving over OP by $+3.3$ pp at 1-shot and by $+8.2$ pp at 16-shot, reaching 76.50% at 16-shot. The geodesic path is marginally better than the linear one (0.43 pp at 16-shot), supporting our choice of spherical interpolation for normalized features. Finally, our *combined* variant is the best at every shot count, with the largest gain at very low K ($+16.2$ pp at 1-shot over our 'independent pair' variant alone) and consistently positive improvements at higher shot counts, further confirming that the two signals are complementary.

Table 4. **Ablation of method components on ImageNet.**

method	setting	1-shot	2-shot	4-shot	8-shot	16-shot
CLIP Zero-Shot	joint encoder (no adaptation)	68.98	68.98	68.98	68.98	68.98
OP only	independent encoder + OP only	53.05	60.27	64.16	67.13	68.25
ours (independent)	independent + OP + geodesic flow	56.33	64.69	70.56	74.20	76.50
ours (combined)	joint-independent combination	**72.51**	**73.95**	**75.99**	**77.97**	**78.95**

D.3. Flow Network (Residual MLP) Architecture

Our velocity network is a residual MLP (SiLU) with time conditioning. We ablate *depth*, *width*, and *time_dim* on ImageNet (16-shot) using **DINOv2-B** for vision and **Qwen-8B** for text, with geodesic flows and OP alignment fixed. Results in Table 5 (Top-1, %) report a 4 value sweep over the depth, width and time_dim parameters, where default values are underlined and are used for the independent sweeping of each other two parameters. It shows small variance ($\leq$0.34 pp) per row, indicating

that our method is stable to reasonable architectural choices. A shallower/wider model yields a slight gain, but our default 4×1536 with time_dim $= 256$ strikes a good accuracy/latency trade-off and is used throughout.

Table 5. **Residual MLP ablation on ImageNet 16-shot (DINOv2-B + Qwen-8B).** Top-1 (%). Best in **bold**; defaults are underlined.

Depth	1	2	4	8
Acc. (%)	**76.47**	**76.47**	76.37	76.17
Width	512	1536	2048	4096
Acc. (%)	75.83	76.37	76.47	**76.73**
time_dim	128	256	512	1024
Acc. (%)	76.30	76.37	76.43	**76.47**

E. Detailed BiomedCoOp Per-Dataset Results

Table 6. **Detailed BiomedCoOp per-dataset results.** Per-dataset accuracies (%) extending Table 1. BiomedCLIP zero-shot accuracy is shown for reference. *Indep.* and *combined* denote our method without and with combination with BiomedCLIP zero-shot logits, respectively. **Bold**/*italics* denote best/second-best per shot count among BiomedCoOp, *indep.*, and *combined*.

dataset	**zero-shot**	**1-shot**			**4-shot**			**16-shot**		
	BiomedCLIP	**BiomedCoOp**	**indep.**	**combined**	**BiomedCoOp**	**indep.**	**combined**	**BiomedCoOp**	**indep.**	**combined**
BUSI	38.14	**50.71**	*39.27*	37.29	**59.32**	*51.69*	47.88	70.34	*74.58*	**75.00**
KneeXray	33.51	**36.13**	28.30	*35.21*	*35.91*	31.82	**38.29**	**39.69**	35.93	*37.14*
CHMNIST	34.38	*59.82*	**64.03**	52.46	71.19	*77.75*	**77.86**	79.05	*87.54*	**88.10**
BTMRI	63.36	**65.08**	53.07	*62.19*	**77.23**	72.56	*74.68*	*83.30*	79.27	**84.07**
COVID_19	62.97	**72.64**	52.41	*67.75*	**73.28**	62.79	*72.10*	**78.72**	70.92	*74.07*
CTKidney	38.29	**56.13**	37.89	*38.66*	**66.50**	52.43	*54.83*	**83.20**	75.73	*78.63*
DermaMNIST	21.24	**58.64**	41.93	*49.33*	**60.07**	*52.95*	50.70	**62.59**	*57.93*	56.69
Kvasir	50.50	**62.17**	53.58	*61.83*	74.08	*75.00*	**80.83**	78.89	*85.50*	**85.92**
LungColon	39.96	77.56	**83.40**	*82.84*	85.60	**90.51**	*89.39*	92.68	**94.80**	*94.17*
OCTMNIST	19.96	**51.83**	*46.15*	26.08	54.73	**60.33**	*56.97*	66.93	*70.95*	**73.28**
RETINA	28.56	36.64	**58.86**	*58.47*	45.58	**64.35**	*62.03*	61.28	**74.39**	*72.47*
average	42.05	**57.03**	50.94	*52.01*	*63.95*	63.17	**64.14**	72.42	*73.71*	**74.41**

F. Computational Cost and Scalability

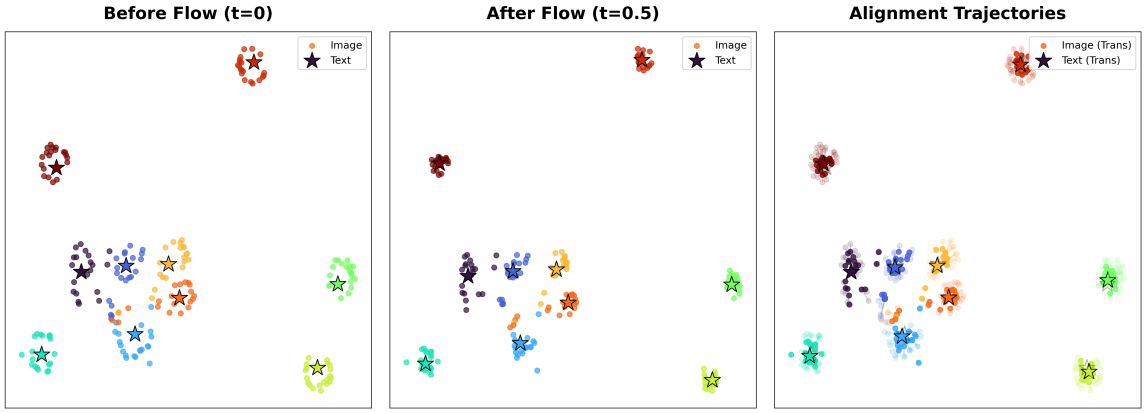
We benchmarked the training and inference efficiency of our method against standard baselines on the 16-shot ImageNet task using a single NVIDIA A100 GPU. As detailed in Table 7, our method demonstrates a superior efficiency profile: (i) **Training Efficiency:** Our method is significantly faster to train than competing adapter methods. Our standard model (Depth=2) reaches convergence in 17 minutes, which is approximately 3× faster than CLIP-Adapter (50 min) and over 60× faster than prompt tuning methods like CoOp (14h 40min); (ii) **Inference Latency:** Despite using an ODE solver, the added latency per query is negligible (∼1.6 ms). The total inference time (11.8 ms) is effectively identical to the baselines; (iii) **Scalability:** Our method offers a flexible trade-off between capacity and cost. Our lightest configuration (Depth=1) already outperforms baselines with only 13 minutes of training, while increasing depth to 4 yields further accuracy gains for a marginal increase in cost.

Table 7. Computational cost comparison for 16-shot ImageNet adaptation (ResNet-50 backbone on a single A100). Our method dominates baselines, achieving SOTA accuracy, 3-4x faster training than CLIP-Adapter, and identical inference speed.

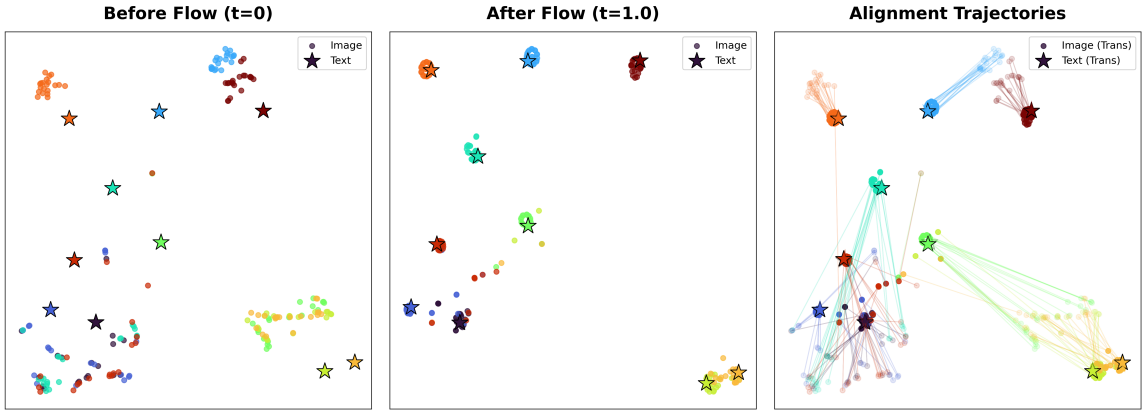
Model	Train Time	Trainable Params	Infer Speed (ms)
Zero-Shot	0	0	10.2
Linear Probe	13min	1.02M	-
CoOp	14h 40min	0.02M	299.6
CLIP-A	50min	0.52M	10.6
ours	15min	6.93M	11.5

We attribute this efficiency to our flow-matching objective, which learns from direct (image, text) pairs, avoiding the computationally expensive batch-wide pairwise similarity calculations required by contrastive adaptation losses.

G. OP Initializations and Flow Trajectories



(a) StanfordCars



(b) FGVCAircraft

Figure 3. **t-SNE of post-OP text–image alignment on two fine-grained datasets.** Text prototypes (stars) are plotted alongside K-shot image features (dots) after Orthogonal Procrustes alignment on the 16-shot support set. Class colors are shared across modalities. On StanfordCars (a) the OP alignment is already strong, each text prototype sits near its image cluster. On FGVCAircraft (b) the modalities remain largely disjoint after OP. The flow adapter closes this residual gap by transporting image features toward the text cluster structure in a continuous, class-conditioned path, enabling accurate classification even when the initial linear alignment is weak. See Sec. G for details. See Sec. G for details.

To qualitatively assess how our method bridges the gap between the two frozen encoders, we visualize their latent spaces via t-SNE. Figure 3 shows the joint image-text embedding for two fine-grained benchmarks: StanfordCars (Krause et al., 2013) and FGVCAircraft (Maji et al., 2013). The setup uses our uni-modal configuration: a DINOv3-B image encoder paired with a CLIP ViT-B/16 text encoder, aligned by Orthogonal Procrustes (OP) fit on the $K=16$ support set.

Each panel overlays class-colored image features (dots) with the corresponding text prototypes (stars) after applying the OP to the text side. We highlight a small, representative subset of classes for readability.

Observations. The two datasets represent different degrees of initial alignment. On StanfordCars (Fig. 3(a)), the OP alignment is already strong: text prototypes land inside or adjacent to their image clusters, and the residual correction needed from the flow adapter is small. On FGVCAircraft (Fig. 3(b)), the linear OP alone is insufficient: image clusters and text prototypes occupy largely disjoint regions, and many classes visibly overlap.

This contrast motivates the flow adapter. Rather than relying on a linear mapping to close the modality gap, our method learns a continuous, class-conditioned transport that moves image features toward the structure induced by the text prototypes. The visualization thus makes concrete the failure mode of linear alignment on harder, fine-grained domains and justifies a non-linear refinement step, consistent with the quantitative gains reported in our main experiments.