Can Experts Adapt Without Training? On Test-Time Modality Generalization in MVLMs

Raza Imam*,[1](✉), Darakshan Rashid*,[3], Yutong Xie[1], Dwarikanath Mahapatra[2], Brejesh Lall[3], and Mohammad Yaqub[1]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Khalifa University, Abu Dhabi, UAE
[3] Indian Institute of Technology Delhi, India

Abstract. Medical vision-language models (MVLMs) promise broad zero-shot generalization, yet their reliability collapses when confronted with unseen modalities and domains, precisely where clinical robustness matters most. To address this gap, we revisit test-time modality generalization from the perspective of Mixture-of-Experts (MoE) and ask: *can experts route-and-adapt without any optimization during inference*? We identify a fundamental *specialization-generalization dilemma* at test time, where blindly aggregating modality experts dilutes modality-specific knowledge, while selecting one highly confident expert risks mismatch under shift. To address this, we propose MoBE: a fully optimization-free framework that performs dynamic expert selection and adaptation at test time. MoBE combines entropy-guided dynamic routing in MoE settings with expert-wise Bayesian adaptation, enabling experts to update their confidence and adapt online without gradient updates. Without parametric updates, MoBE augments a static MVLM with test-time routing and online statistics, achieving average accuracy gains of +4.72, +7.17, and +4.3 over state-of-the-art TTA methods across seen, unseen, and heterogeneous medical benchmarks, highlighting the effectiveness of training-free expert adaptation for robust modality generalization.

Keywords: Test-Time Adaptation · Multimodal · Mixture-of-Experts.

1 Introduction

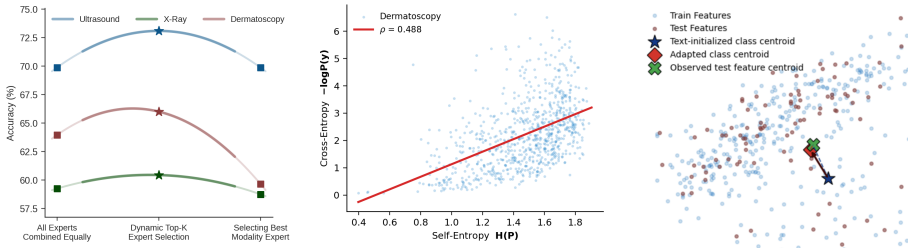
Recent medical vision-language models (MVLMs) enable strong zero-shot transfer across imaging tasks and modalities, but this performance often assumes test data matches training conditions [7,6]. In clinical deployment, distribution shifts from unseen modalities, scanners, and acquisition protocols are common, and accuracy can degrade substantially [1,11,9,3].

Expert Specialization for Modality Heterogeneity: To address modality heterogeneity, prior work [22] has explored expert-based architectures that

Dataset and Code is available at: Github

Corresponding Author: Raza Imam ✉(raza.imam@mbzuai.ac.ae)

Equal Contribution *



(a) Identified Dilemma (b) Pearson Correlation (c) Centroid Misalignment

Fig. 1: (a) The specialization-generalization dilemma induced by uniform expert fusion v/s single-expert reliance at test time. (b) Routing by self-entropy closely matches oracle cross-entropy routing, as they positively correlate. (c) Despite correct routing, pretrained class centroids can misalign with test features, motivating online alignment.

partition representation learning across modalities. Mixture-of-Experts (MoE) models offer a natural mechanism for leveraging modality-specific knowledge [23], but their effectiveness at test time critically depends on how experts are selected and combined under distribution shift. This raises a central question for test-time modality generalization: *how should expert specialization be exploited when the reliability of each expert is unknown at inference?*

Limitations of Existing Adaptation: MoE-based MVLMs often use fixed fusion or optimize routing at test time. For example, MoME [22] adapts routing parameters via test-time optimization, aligning with gradient-based adaptation, e.g., TNT, TPT, and TTL [5,19,4]. In contrast, cache-based methods (e.g., TDA, BoostAdapter) avoid optimization via test-time memory [8,25], but operate on a single model and do not address expert selection under modality shifts.

The Specialization-Generalization Dilemma: These observations reveal an underexplored dilemma in the generalization of test-time modality: the *specialization-generalization dilemma*, where blindly aggregating experts dilutes modality-specific knowledge, while committing to a single expert risks severe mismatch under shift [26,13]. Addressing this dilemma, therefore, requires *estimating expert reliability on-the-fly at inference time*, ideally in a training-free manner that avoids backpropagation (Fig. 1a). This motivates test-time adaptation in mixture-of-modality-experts settings, where dynamic expert selection and lightweight adaptation can be performed online.

Towards Training-Free Expert Adaptation: In this work, we propose an optimization-free framework that integrates dynamic expert routing with expert-wise Bayesian adaptation at test time. To ensure unsupervised routing matches supervised performance [15,20], Fig. 1b empirically validates entropy as a cross-entropy proxy for routing. Also, instead of updating model parameters, each expert maintains an online posterior, accumulated across the test stream, that captures its evolving confidence and specialization, guiding both expert selection and prediction aggregation, enabling adaptation to modality shifts (Fig. 1c).

Contributions: Our contributions are threefold: **(1)** We reveal the *specialization-generalization dilemma* that cripples test-time modality adaptation in Medical VLMs; **(2)** We present MoBE: a fully *optimization-free* inference framework uniting dynamic expert routing with per-expert Bayesian adaptation; and **(3)** We demonstrate *consistent state-of-the-art gains* across standard and heterogeneous medical benchmarks, proving training-free expert specialization enables robust modality generalization.

2 Methodology

Our modality-generalizable test-time adaptation is a two-stage process: first, dynamic-k routing addresses the *specialization-generalization dilemma* by selecting contextually reliable modality experts through predictive uncertainty; second, Expert Bayesian Adaptation (EBA) refines the predictions of selected experts through online posterior updates via exponential moving averages (Fig. 2). This yields an optimization-free, MoE-based TTA system that continuously adapts how fixed representations are combined and refined, achieving robust performance without gradient updates or source data access.

2.1 Preliminaries: Modality-Specialized MVLM Backbone

Given a sequence of test images $\{x_i\}_{i=1}^{n}$ that arrive sequentially from an unseen modality, the model is required to predict the label for each image x_i immediately in an online manner. For each incoming image, the frozen image encoder f_θ of a multimodal vision-language model (MVLM) (e.g., BiomedCLIP [24]) extracts a visual representation $h_i = f_\theta(x_i) \in \mathbb{R}^D$. Zero-shot classification is performed by computing the similarity between the visual feature h_i and text-encoder embeddings $w_c \in \mathbb{R}^D$, corresponding to class prompts of the form "a [class] of [modality]", yielding logits $s_{i,c} = h_i^\top w_c$.

To leverage modality-specific knowledge, following MoME [22], we attach E modality-specialized MLP experts $\{g_e\}_{e=1}^{E}$ in parallel to the frozen image encoder's classification head. Each $g_e : \mathbb{R}^D \to \mathbb{R}^C$ maps visual features to class logits tailored to modality m_e, producing $s_e = g_e(h)$ and softmax probabilities $p_e = \mathrm{softmax}(s_e)$. The challenge is selecting and combining these parallel experts when the test modality is unknown.

2.2 MoBE: Mixture of Bayesian Experts in Medical Imaging

A. Dynamic-k Entropy-Guided Expert Routing: As shown in Fig. 1a, both uniform expert fusion and single-expert selection suffer performance degradation by either diluting modality knowledge or risking catastrophic mismatch. We resolve this through dynamic-k routing that adaptively gates experts based on relative predictive uncertainty.

For each test sample, we compute predictive entropy $H(p_e) = -\sum_c p_e(c) \log p_e(c)$ across all E experts, where lower $H(p_e)$ signals higher modality alignment.

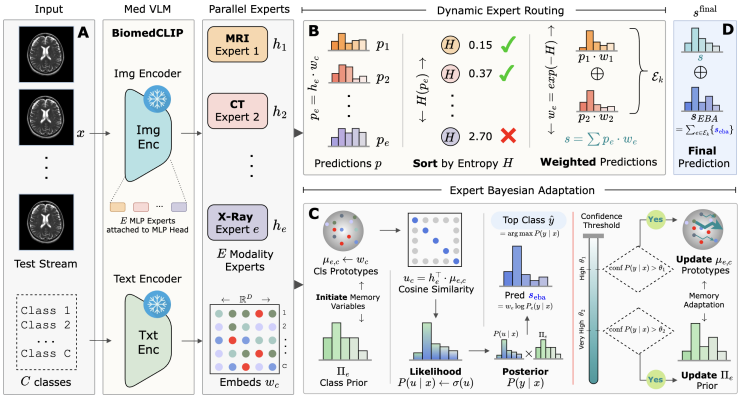


Fig. 2: **MoBE Test-Time Adaptation.** (A) Frozen BiomedCLIP encoder processes an incoming stream of test images. (B) Dynamic-k entropy routing selects the most reliable modality experts from parallel MLP heads. (C) Selected experts are inverse-entropy weighted and refined by EBA through confidence-gated prototype and prior updates using the test stream. (D) Final predictions blend routed logits with Bayesian-adapted outputs.

Rather than fixed top-k, we apply relative thresholding: sort by ascending entropy $H_1 \leq H_2 \leq \cdots \leq H_E$, then select k experts as:

$$\mathcal{E}_k = \{e_j \mid H_j - H_1 < \tau\}, \quad k = |\mathcal{E}_k|, \quad j \in 1, \ldots, E, \tag{1}$$

where τ is a modality-agnostic gap threshold. This way, we select and gate experts who are nearly as confident as the most confident expert, and exclude clearly mismatched ones. A minimum of two experts is enforced when $E > 1$, maintaining expert diversity and inference efficiency [22].

Following observations from TTL and DeYo [4,12], we note that inverse-entropy weighting produces mixing coefficients that better capture relative confidence differences than softmax:

$$w_e \propto \frac{1}{\exp(H(p_e))}, \quad e \in \mathcal{E}_k. \tag{2}$$

Gated logits become the weighted ensemble $s = \sum_{e \in \mathcal{E}_k} w_e s_e$, where the gating function naturally emphasizes confident modality experts.

B. Expert Bayesian Adaptation (EBA): Once top experts $\mathcal{E}_k$ are selected via dynamic routing, the next step is to adapt their predictions online as the test stream evolves under a distribution shift. Motivated by statistical methods for natural images [2,29], which frame predictions as a factorization of likelihood $P(x \mid y)$ and prior $P(y)$, we observe that existing approaches primarily adapt likelihoods while neglecting prior shifts. To address this, our Expert Bayesian Adaptation (EBA) module maintains per-expert Bayesian posteriors and jointly adapts both factors online, without updating any backbone or expert parameters.

For each selected expert $e \in \mathcal{E}_k$, EBA maintains two persistent *memory variables* across the test stream: (i) class prototypes $\mu_{e,c} \in \mathbb{R}^D$ (one per class c, serving as a moving representative feature/centroid), and (ii) a class-prior vector $\Pi_e \in \Delta^{C-1}$ (a length-C distribution capturing the expert's historical tendency to predict each class). As additional test samples are observed, these accumulated statistics become increasingly accurate summaries of the underlying test distribution.

Initialization: Prototypes are initialized from pretrained text embeddings $\{w_c\}_{c=1}^{C}$ as semantically grounded class anchors: $\mu_{e,c} \leftarrow w_c/\|w_c\|_2$, where $c = 1, \ldots, C$. The class prior is initialized uniformly: $\Pi_e \leftarrow [1/C, 1/C, \ldots, 1/C]$. We also initialize per-expert, per-class update counters $c_{1,e,c}, c_{2,e,c} \in \mathbb{N}$; these count confident prototype and prior updates, respectively, and act as running-average denominators.

Prediction: Given a test feature vector h, we score each class by its similarity to the corresponding prototype: $u_c = \tau_{\text{EBA}} \cdot h^\top \mu_{e,c}$, and convert scores into a probability vector $P(u \mid x) = \text{softmax}(u)$, which we treat as a likelihood-like quantity indicating how well the input matches each class. To incorporate the expert's stored history to attain full *posterior*, we reweight these probabilities by the prior Π_e (elementwise) and renormalize:

$$\tilde{P}(y \mid x) = P(u \mid x) \odot \Pi_e, \qquad P(y \mid x) = \frac{\tilde{P}(y \mid x)}{\sum_c \tilde{P}(y = c \mid x)}. \tag{3}$$

Intuitively, this combines *current similarity to prototypes* with *what this expert has been seeing recently.*

Confidence-gated adaptation: EBA updates its memory only when it is confident, using confidence-gated running averages. (i) *Prototype update (threshold θ_1):* For high-confidence predictions (i.e., if $\max_c P(y = c \mid x) > \theta_1$), we update only the prototype of the predicted class $\hat{y} = \arg\max P(y \mid x)$ by averaging it with the new feature h:

$$\mu_{e,\hat{y}} \leftarrow \frac{c_{1,e,\hat{y}}\, \mu_{e,\hat{y}} + h}{c_{1,e,\hat{y}} + 1}, \qquad c_{1,e,\hat{y}} \leftarrow c_{1,e,\hat{y}} + 1, \tag{4}$$

followed by re-normalizing prototypes to unit length. (ii) *Prior update (higher threshold θ_2):* When confidence exceeds a higher threshold (i.e., if $\max_c P(y = c \mid x) > \theta_2$), we additionally update the class prior using the current posterior $P(y \mid x)$:

$$\Pi_e \leftarrow \frac{c_{2,e,\hat{y}}\, \Pi_e + P(y \mid x)}{c_{2,e,\hat{y}} + 1}, \qquad c_{2,e,\hat{y}} \leftarrow c_{2,e,\hat{y}} + 1. \tag{5}$$

As $c_{1,e,c}$ and $c_{2,e,c}$ increase, each update receives less weight, making $\mu_{e,c}$ and Π_e progressively more stable. This dual adaptation mechanism: (i) moving prototypes toward the observed test feature manifold, and (ii) adjusting class priors based on a history of highly confident posteriors; aligning to the target distribution while preserving pretrained semantic structure. EBA is applied independently to each $e \in \mathcal{E}_k$, with aggregated logits as: $s_{\text{EBA}} = \sum_{e \in \mathcal{E}_k} w_e \log P_e(y \mid x)$.

Final predictions fuse entropy-routed logits with EBA-adapted logits: $s^{\text{final}} = (1-\lambda)\, s + \lambda\, s_{\text{EBA}}$, where $\lambda \in [0,1]$ controls the strength of adaptation.
C. Training & Test-Time Inference: Each modality-specific MLP g_e trains independently on source data from modality m_e, minimizing cross-entropy using the frozen image encoder f_θ. No routing or cross-modal mixing occurs during training. *Test-Time Inference:* The image-text encoders and experts remain frozen. Per test batch: (1) extract h and per-expert logits $\{s_e\}$; (2) compute entropies $\{H(p_e)\}$ and gate dynamic-k set $\mathcal{E}_k$; (3) apply inverse-entropy weighting to produce s; (4) update EBA posteriors for selected experts; (5) blend via s^{final}.

3 Experiments and Results

3.1 Experimental Settings

Benchmarks: Following prior work on modality-level TTA [22], we evaluate on a combined MedMNIST+Med-VTAB benchmark [14], covering 11 datasets across 9 imaging modalities (e.g., X-ray, CT, pathology, microscopy, ultrasound, OCT, and MRI). Med-VTAB adds MRI to mitigate modality gaps in MedMNIST and enables controlled assessment of modality generalization. Furthermore, we evaluate on a heterogeneous few-shot suite [10] with 11 datasets spanning 10 organs and 9 modalities (including MRI, ultrasound, endoscopy, and histopathology), to test robustness under more diverse clinical imaging conditions.
Implementation Details: We use the pretrained BiomedCLIP ViT-B/16 backbone [24] with frozen image and text encoders. We attach E=5 modality-specialized MLP experts in parallel to the image encoder's classification head. Similar to [22], experts are derived from ROCOV2 [18] by partitioning data into five modality groups (CT, MRI, X-ray, ultrasound, angiogram) and fine-tuning only the vision-side MLP layers per modality; the resulting MLP weights are extracted as fixed expert branches in MoBE. These modality experts have the same architecture as in BiomedCLIP's MLP. All images are processed using the standard BiomedCLIP preprocessing with no additional augmentation. We evaluate with batch size 1 on a single NVIDIA A6000 (48GB) and report Top-1 accuracy. Dataset-specific hyperparameter values for MoBE are available at Github.

Table 1: Performance of different VLMs and TTA methods on MedMNIST and Med-VTAB datasets [22] (% Accuracy).

Method	ChestMNIST	MedVTAB	BreastMNIST	OrganAMNIST	OrganCMNIST	OrganSMNIST	Average
Modality	X-Ray	MRI	Ultrasound	CT	CT	CT	Accuracy
BiomedCLIP	53.29	51.77	32.69	25.47	17.60	17.15	33.00
CoOp [52]	53.30	52.71	33.58	26.04	17.65	17.59	33.48
CoCoOp [51]	53.36	52.75	33.51	26.19	17.88	17.71	33.57
TENT [37]	41.81	45.86	30.15	15.47	9.61	10.02	25.49
EATA [30]	44.63	51.98	30.83	22.31	15.82	15.55	30.19
TPT [36]	53.33	53.22	35.77	26.12	17.67	17.60	33.95
WATT [31]	53.42	56.10	35.97	25.89	16.84	16.77	34.17
TDA [18]	53.50	55.69	45.87	26.40	17.91	17.82	36.20
MoME	54.72	58.12	52.56	**29.19**	18.45	19.12	38.69
MoBE (Ours)	**60.44**	**60.72**	**73.08**	24.37	**19.84**	**21.99**	**43.41**

Table 2: Performance of different VLMs and TTA methods on MedMNIST datasets with unseen modalities [22] (% Accuracy).

Method	PathMNIST	DermaMNIST	OCTMNIST	BloodMNIST	TissueMNIST	Average
Modality	Colon Pathology	Dermatoscope	Retinal OCT	Microscope	Microscope	Accuracy
BiomedCLIP [49]	28.31	17.71	27.70	23.82	4.89	20.49
CoOp [52]	28.59	18.07	29.98	24.30	4.95	21.18
CoCoOp [51]	28.50	18.15	30.82	24.36	5.10	21.39
TENT [37]	20.48	14.05	21.96	15.54	3.22	15.05
EATA [30]	24.72	14.20	27.82	19.35	4.60	18.14
TPT [36]	28.34	20.03	31.57	25.14	5.07	22.03
WATT [31]	28.45	29.64	33.50	25.12	4.91	24.32
TDA [18]	28.21	23.18	31.65	24.79	4.96	22.56
MOME	29.74	57.51	36.20	**26.31**	5.18	30.99
MoBE	**48.69**	**66.00**	**54.20**	15.70	**6.22**	**38.16**

Table 3: Performance of different VLMs and TTA methods on heterogeneous medical datasets [10] (% Accuracy).

Method	BTMRI	COVID-QU-Ex	CTKIDNEY	DermaMNIST	Kvasir	CHMNIST
Modality	MRI	X-Ray	CT	Dermatoscopy	Endoscopy	Histopathology
BiomedCLIP [49]	51.77	68.72	49.01	17.71	47.25	23.67
TDA [18]	55.69	64.45	54.71	23.18	**52.08**	**27.53**
MoBE	**62.90**	**69.86**	**55.72**	**66.00**	50.42	24.87

Method	LC25000	RETINA	KneeXray	OCTMNIST	BUSI	Average
Modality	Histopathology	Fundus Photography	X-Ray	OCT	Ultrasound	Accuracy
BiomedCLIP [49]	33.87	23.74	**40.28**	27.70	41.10	39.24
TDA [18]	**40.80**	27.60	38.47	31.65	45.34	42.19
MoBE	37.68	**29.73**	40.10	**54.20**	**45.76**	**46.49**

3.2 Comparison with State-of-the-art

We benchmark MoBE against prominent adaptation methods on BiomedCLIP, including TENT [21], EATA [16], CoOp [27], CoCoOp [28], TPT [19], WATT [17], and TDA [16]. While CoOp and CoCoOp are prompt-learning methods rather than strict TTA, we include them following common protocols in [22].

Results on MedMNIST and MedVTAB (seen modalities): Tab. 1 reports datasets whose modalities are covered by our modality-specialized experts. MoBE achieves the best average accuracy (43.41%), improving over BiomedCLIP (33.00%) and MoME (38.69%). Gains are consistent on ChestMNIST, MedVTAB, and BreastMNIST (notably 73.08% on BreastMNIST), and also on OrganCMNIST/OrganSMNIST. OrganAMNIST is the main exception, likely due to weaker modality cues and heterogeneous class semantics that make entropy-based expert selection less reliable.

Results on MedMNIST with unseen modalities: Tab. 2 evaluates generalization to modalities unseen during expert pre-training. MoBE substantially improves the average (38.16%) over BiomedCLIP (20.49%), TDA (22.56%), and MoME (30.99%), with strong gains on PathMNIST, DermaMNIST, OCTMNIST, and TissueMNIST. BloodMNIST remains challenging, plausibly because microscopy shifts are not well covered by our five expert modalities, reducing routing discriminability.

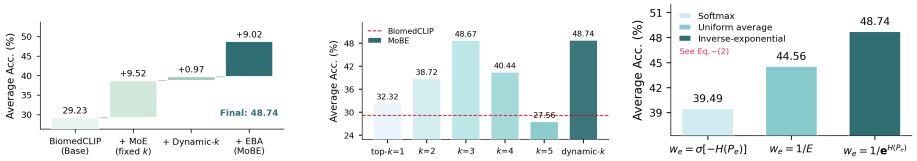
Table 4: Performance comparison with different frozen backbones.

Backbone ↓	Seen-Avg	Unseen-Avg	Average
PubmedCLIP	35.89	10.37	23.13
+**MoBE** (Ours)	**48.13**	**10.98**	**29.56**
BiomedCLIP	35.45	23.01	29.23
+**MoBE** (Ours)	**40.14**	**57.35**	**48.74**

Table 5: Computation analysis on X-ray modality (ChestMNIST, ∼22k samples).

Method ↓	No Backprop?	Infer-Time	Speed*	Accuracy	ΔGain
BiomedCLIP	✔	2m30s	144 i/s	53.29	0.00
MoME	✘	12m57s	28 i/s	54.72	1.43
MoBE (Ours)	✔	**11m10s**	**35 i/s**	**60.44**	**7.15**

*i/s denotes inference speed in images-per-second



(a) Component-wise result (b) Fixed-k *vs.* dynamic-k (c) Expert weighting effect

Fig. 3: (a) **Component-wise performance:** Each component of MoBE improves generalization. (b) **Expert selection effect:** MoBE's dynamic-k routing matches the best fixed top-k while remaining fully automatic. (c) **Expert weighting effect:** MoBE's inverse-exponential entropy outperforms averaging and softmax as gating functions.

Results on Heterogeneous Dataset: Tab. 3 extends evaluation to heterogeneous datasets (e.g., endoscopy, fundus photography, histopathology). Since MoME is not reproducible here due to code unavailability, we compare BiomedCLIP, TDA, and MoBE. Averaged over 11 datasets and 9 modalities, MoBE achieves the best average (46.49%) vs. TDA (42.19%), with large gains on OCTMNIST and competitive results on RETINA/BUSI. TDA is slightly stronger on some sets (e.g., Kvasir/CHMNIST), where style-driven shifts may favor cache-based likelihood adaptation over expert selection.

3.3 Ablation Studies

We ablate MoBE components by averaging performance over *seen* modalities (ChestMNIST, OrganCMNIST) and *unseen* modalities (PathMNIST, DermaMNIST), assessing the impact of different design choices.

Component-wise Ablation: Relative to frozen BiomedCLIP, MoE with fixed top-k improves accuracy; dynamic-k further improves while eliminating the need to tune k. EBA yields the largest gain via online refinement of per-expert posteriors, and combined routing-and-adaptation performs best (Fig. 3a).

Fixed Top-k vs. Adaptive-k: Fixed top-k is sensitive to k (too many experts can hurt). Dynamic-k matches the best fixed-k performance while remaining fully automatic, expanding the expert set only when additional experts have comparable uncertainty (Fig. 3b).

Expert Weighting Mechanism: For gating the selected experts, uniform averaging improves over softmax weighting, while inverse-exponential entropy weighting performs best, emphasizing low-entropy experts (Fig. 3c).

Compute Analysis: Compared to optimization-based MoME, MoBE is inference-only (no backprop), achieving higher accuracy with substantially lower overhead. The cost comes from extra forward passes over multiple experts, making it slower than single-model BiomedCLIP inference (Tab. 5).
Effect of Different Backbones: Across backbones (e.g., PubmedCLIP vs. BiomedCLIP), MoBE consistently improves average accuracy, with larger gains on unseen modalities, indicating robust backbone transferability (Tab. 4).

4 Conclusion

We address test-time modality generalization in MVLMs with our proposed MoBE: a test-time route-and-adapt MoE framework combining adaptive expert selection and Bayesian adaptation. MoBE consistently improves performance across seen, unseen, and heterogeneous benchmarks across backbones. However, MoBE incurs higher inference cost than frozen BiomedCLIP due to multiple expert forward passes. Future work should reduce this overhead (e.g., via early-exit routing, pruning/distillation, or caching) and improve robustness when available expert-bank coverage of the test distribution is limited.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cheng, Z., Ong, A.Y., Wagner, S.K., Merle, D.A., Ju, L., Zhang, H., Chen, R., Pang, L., Li, B., He, T., et al.: Understanding the robustness of vision-language models to medical image artefacts. NPJ Digital Medicine **8**(1), 727 (2025) 1
2. Han, Z., Yang, J., Wang, G., Li, J., Xu, Q., Shou, M.Z., Zhang, C.: Dota: Distributional test-time adaptation of vision-language models. arXiv preprint arXiv:2409.19375 (2024) 4
3. Hanif, A., Shamshad, F., Awais, M., Naseer, M., Khan, F.S., Nandakumar, K., Khan, S., Anwer, R.M.: Baple: Backdoor attacks on medical foundational models using prompt learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 443–453. Springer (2024) 1
4. Imam, R., Gani, H., Huzaifa, M., Nandakumar, K.: Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5449–5459. IEEE (2025) 2, 4
5. Imam, R., Hanif, A., Zhang, J., Dawoud, K.W., Kementchedjhieva, Y., Yaqub, M.: Noise is an efficient learner for zero-shot vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5820–5829 (2025) 2
6. Imam, R., Marew, R., Yaqub, M.: On the robustness of medical vision-language models: Are they truly generalizable? In: Annual Conference on Medical Image Understanding and Analysis. pp. 233–256. Springer (2025) 1

7. Ju, L., Zhou, S., Zhou, Y., Lu, H., Zhu, Z., Keane, P.A., Ge, Z.: Delving into out-of-distribution detection with medical vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 133–143. Springer (2025) 1
8. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14162–14171 (2024) 2
9. Khan, U., Nawaz, U., Sheikh, T.T., Hanif, A., Yaqub, M.: Guardian: Guarding against uncertainty and adversarial risks in robot-assisted surgeries. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 59–69. Springer (2024) 1
10. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Biomedcoop: Learning to prompt for biomedical vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14766–14776 (2025) 6, 7
11. Korevaar, S., Tennakoon, R., Bab-Hadiashar, A.: Failure to achieve domain invariance with domain generalization algorithms: An analysis in medical imaging. IEEE Access **11**, 39351–39372 (2023) 1
12. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. arXiv preprint arXiv:2403.07366 (2024) 4
13. Li, B., Shen, Y., Yang, J., Wang, Y., Ren, J., Che, T., Zhang, J., Liu, Z.: Sparse mixture-of-experts are domain generalizable learners. arXiv preprint arXiv:2206.04046 (2022) 2
14. Mo, S., Luo, X., Wang, Y., Li, D.: A large-scale medical visual task adaptation benchmark. arXiv preprint arXiv:2404.12876 (2024) 6
15. Nikolic, S., Oguz, I., Psaltis, D.: Exploring expert specialization through unsupervised training in sparse mixture of experts. arXiv preprint arXiv:2509.10025 (2025) 2
16. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022) 7
17. Osowiechi, D., Noori, M., Vargas Hakim, G., Yazdanpanah, M., Bahri, A., Cheraghalikhani, M., Dastani, S., Beizaee, F., Ayed, I., Desrosiers, C.: Watt: Weight average test time adaptation of clip. Advances in neural information processing systems **37**, 48015–48044 (2024) 7
18. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Koitka, S., Pelka, O., Abacha, A.B., G. Seco de Herrera, A., et al.: Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. Scientific Data **11**(1), 688 (2024) 6
19. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022) 2, 7
20. Su, Y., Liu, Y.: Variational inference, entropy, and orthogonality: A unified theory of mixture-of-experts. arXiv preprint arXiv:2601.03577 (2026) 2
21. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020) 7
22. Wang, H., Yu, Y., Zheng, H., Zhang, T.: Test-time adaptation of medical vision-language models with mixture of modality experts. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4649–4658 (2025) 1, 2, 3, 4, 6, 7

23. Wang, X., Yang, C.: Moe-health: A mixture of experts framework for robust multimodal healthcare prediction. In: Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 1–9 (2025) 2
24. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023) 3, 6
25. Zhang, T., Wang, J., Guo, H., Dai, T., Chen, B., Xia, S.T.: Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. Advances in Neural Information Processing Systems **37**, 67795–67825 (2024) 2
26. Zhong, T., Chi, Z., Gu, L., Wang, Y., Yu, Y., Tang, J.: Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. Advances in Neural Information Processing Systems **35**, 22243–22257 (2022) 2
27. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022) 7
28. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022) 7
29. Zhou, L., Ye, M., Li, S., Li, N., Zhu, X., Deng, L., Liu, H., Lei, Z.: Bayesian test-time adaptation for vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29999–30009 (2025) 4