Medical Cross-Modal Prompt Hashing with Robust Noisy Correspondence Learning

Yishu Liu[1][0009-0004-7924-4039], Zhongqi Wu[2][0009-0003-3421-5528],
Bingzhi Chen[✉,3][0000-0002-2497-6214], Zheng Zhang[1][0000-0003-1470-6998], and
Guangming Lu[✉,1][0000-0003-1578-2634]

[1] Harbin Institute of Technology, Shenzhen, China
[2] South China Normal University, Guangzhou, China
[3] Beijing Institute of Technology, Zhuhai, China
chenbingzhi.smile@gmail.com, luguangm@hit.edu.cn **

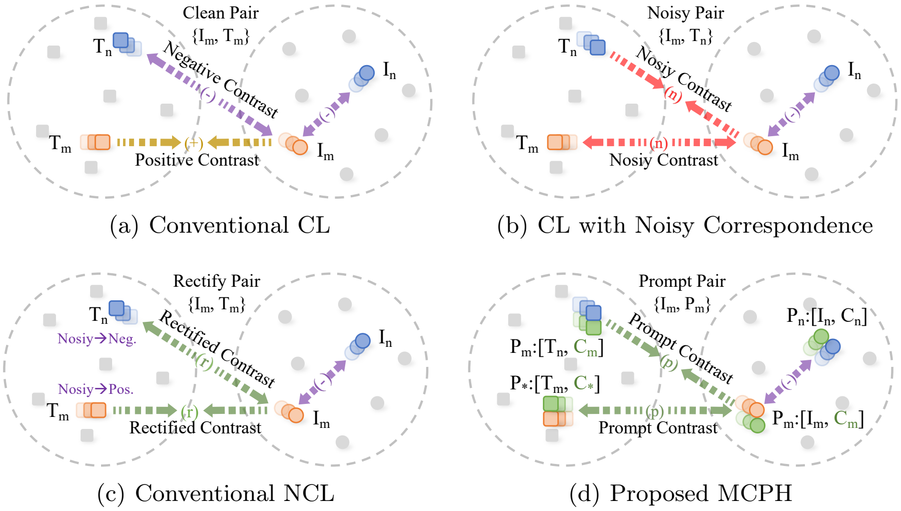
Abstract. In the realm of medical data analysis, medical cross-modal hashing (Med-CMH) has emerged as a promising approach to facilitate fast similarity search across multi-modal medical data. However, due to human subjective deviation or semantic ambiguity, the presence of *noisy correspondence* across medical modalities exacerbates the challenge of the heterogeneous gap in cross-modal learning. To eliminate clinical noisy correspondence, this paper proposes a novel medical cross-modal prompt hashing (MCPH) that incorporates multi-modal prompt optimization with noise-robust contrastive constraint for facilitating noisy correspondence issues. Benefitting from the robust reasoning capabilities inherent in medical large-scale models, we design a visual-textual prompt learning paradigm to collaboratively enhance alignment and contextual awareness between the medical visual and textual representations. By providing targeted prompts and cues from the medical large language model (LLM), i.e., *CheXagent*, multi-modal prompt learning facilitates the extraction of relevant features and associations, empowering the model with actionable insights and decision support. Furthermore, a noise-robust contrastive learning strategy is dedicated to dynamically adjusting the intensity of contrastive learning across modalities, thereby enhancing the contrast strength of positive pairs while mitigating the influence of noisy correspondence pairs. Extensive experiments on multiple benchmark datasets demonstrate that our MCPH surpasses the state-of-the-art baselines.

Keywords: Medical Cross-Modal Hashing · Noisy Correspondence Learning · Large Language Model · Prompt Learning · Contrastive Learning

1 Introduction

Driven by the gradual improvement of medical information storage standards, multi-modal medical data have exhibited an unprecedentedly explosive growth. In recent years, the escalating demand has emerged for methodologies aimed at diminishing storage necessities while enabling rapid and accurate medical data

** ✉ Corresponding authors: Bingzhi Chen and Guangming Lu.



(a) Conventional CL (b) CL with Noisy Correspondence

(c) Conventional NCL (d) Proposed MCPH

Fig. 1. Illustrations of (a) Conventional Contrastive Learning (CL), (b) CL with Noisy Correspondence, (c) Conventional Noisy Correspondence Learning (NCL), and (d) The proposed MCPH. Instead of selecting and rectifying the noisy correspondence, our MCPH approach effectively incorporates the multi-modal prompt optimization with noisy-robust contrastive constraint to further enhance noisy correspondence learning.

retrieval. Therefore, the exploration of medical cross-modal hashing (Med-CMH) [13,32] holds substantial research value and clinical relevance in the medical domain. In clinical settings, Med-CMH aims to retrieve semantically similar medical instances across different modalities. Technically, Med-CMH can provide previously similar images and corresponding treatment records to physicians, thereby enhancing the efficiency and accuracy of decision-making.

Typically, the primary challenges of Med-CMH involve bridging the heterogeneity gap across different modalities [17,25]. Extensive deep cross-modal hashing (CMH) methods [1,28,22] have been developed to facilitate fast similarity search across different modalities, ranging from natural scenes to medical settings. Despite achieving progress in Med-CMH [27,30,35], their success relies on an implicit assumption, i.e., **the semantic correspondence across modalities of training data is ideally correct and noise-free**. *Due to various factors such as inconsistent data sources, human subjective deviation, and semantic ambiguity, it is inevitable to encounter mismatched pairs erroneously considered as matches during the medical data collection process [9,2,33], termed as "**noisy correspondence**".* As shown in Fig. 1, conventional NCL studies focus on dividing the training data into clean and noisy subsets, followed by rectifying the noisy correspondence labels. That is, the key challenge lies in estimating and selecting accurate soft correspondence labels for these noisy data pairs. However, the effectiveness of these methods depends on their respective noise pair selection algorithms, which may inadvertently introduce additional noise.

In this paper, we propose a medical cross-modal prompt hashing (MCPH) framework, which dexterously incorporates multi-modal prompt optimization with noisy-robust contrastive constraint to enhance noisy correspondence learning. In comparison to the state-of-the-art methods [9,7,33], the proposed MCPH method mainly benefits from two well-established prompt mechanisms, i.e., visual prompt learning (VPL) and textual prompt learning (TPL), which are designed to enhance alignment and contextual awareness between the medical visual and textual representations. Benefiting from the robust reasoning capabilities inherent in medical large language models (LLM) such as CheXagent, multi-modal prompt learning enables the extraction of pertinent features and associations, thereby empowering the model with actionable insights and decision support. Furthermore, we propose a noise-robust contrastive learning strategy to dynamically adjust the intensity of contrastive learning across modalities. This strategy aims to enhance the contrast strength of positive pairs while mitigating the influence of noisy correspondence pairs. Our proposed MCPH framework is comprehensively evaluated on multiple large-scale datasets, i.e., Open-I and MIMIC-CXR. The promising performance collectively demonstrates the effectiveness and superiority of our MCPH over state-of-the-art algorithms.

2 Related Work

2.1 Cross-modal Hashing&Retrieval

With the powerful representation ability and privileged efficiency, CMH has been extensively researched and typically categorized into two main types: shallow CMH methods [34,19,31] and deep CMH methods [10,15,1,28,22,29]. Correspondingly, shallow CMH methods always employ a two-stage learning paradigm involving feature extraction and hash code learning, while deep CMH methods leverage an end-to-end training manner to generate informative deep features. For example, DAPH [29] generates higher-quality proxy hash codes to improve the retrieval performance. DCHMT [28] explores two transformer encoders to encode images and texts to enhance correlation modeling. MITH [22] leverages both intra- and inter-modal similarity preservation from a multi-granularity perspective within a unified transformer-based framework. Nevertheless, the above works overlook the challenge of noisy correspondence in medical multi-modal data, which can significantly degrade the performance of Med-CMH tasks.

2.2 Noisy Correspondence Learning

The noisy correspondence problem refers to mismatched pairs incorrectly identified as positive ones, which can reduce the generalization of deep learning models. Initially, Huang et al. [9] reveal this emerging challenging direction and propose the NCR method to correct the soft correspondence label. DECL [23] presents an uncertainty-based approach to model the uncertainty of cross-modal correspondence to predict the correct correspondence of paired data. MSCN [7] explores the meta-learning paradigm to provide reliable similarity scores. Furthermore, BiCro [33] rectifies the noisy correspondence labels by soft correspondence label estimation giving only noisily-collected data pairs. Instead of recalibrating the noisy correspondence, our MCPH aims to enhance Med-CMH by effectively leveraging the robust reasoning capabilities inherent in medical LLM by the multi-modal prompt optimation with noise-robust contrastive constraint.

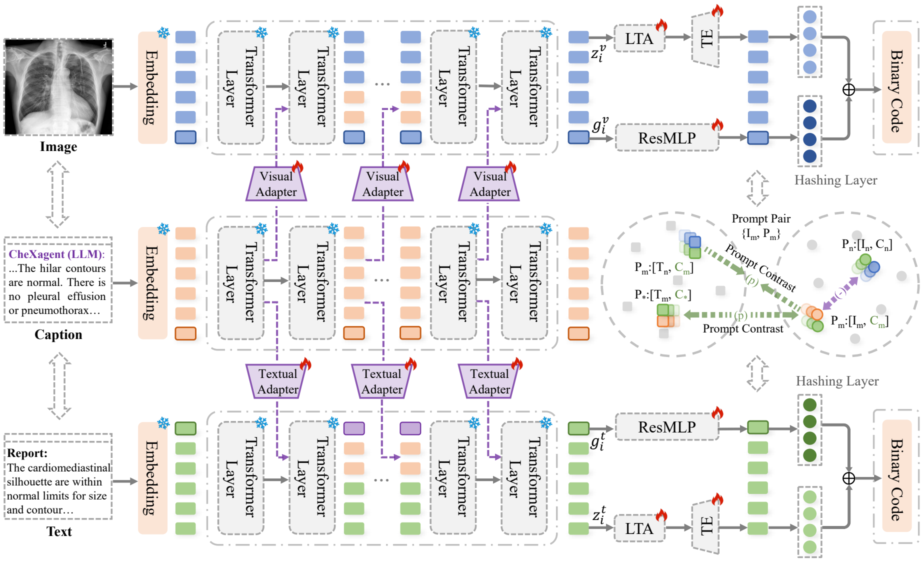


Fig. 2. Illustration of the proposed Medical Cross-Modal Prompt Hashing (MCPH) framework with robust noisy correspondence learning for the task of Med-CMH. Generally, the proposed MCPH is composed of four components, i.e., cross-modal feature learning, visual-textual prompt learning, visual-textual prompt learning, and visual-textual prompt learning.

2.3 Prompt Learning with LLM

Prompt learning [14,16,20,26,3] mainly involves integrating additional tokens such as handcrafted instructions or learnable prompts to fine-tune pre-trained models for downstream tasks using LLM. This strategy eliminates the necessity for substantial modifications to the model's parameters, thus avoiding catastrophic forgetting [8]. For instance, CoOp [37] is the earliest work to employ trainable text prompt vectors for few-shot transferring based on CLIP. CoCoOp [36] underscores the inferior performance of CoOp on novel classes and proposed conditional context optimization on image instances. Recent studies such as DCP [21] and MaPLe [12] further promote strong coupling between the vision-language prompts. However, the effectiveness of multi-modal prompts in the domain of Med-CMH remains to be under-explored and less investigated.

3 Methodology

3.1 Problem Definition

Given a dataset $\mathcal{D}_i = \{(\mathcal{I}_i, \mathcal{T}_i; \mathcal{C}_i; l_i)\}_{i=1}^{N}$, where N indicates the number of training samples, $(\mathcal{I}_i, \mathcal{T}_i)$ is the i-th medical image text pair, $\mathcal{C}_i$ denotes the image captions generated offline by *CheXagent* [4], and l_i indicates the associated label matrix with $\mathcal{Q}$ categories. In practice, it is unavoidable to introduce noisy correspondence into the training data, which renders supervision information unreliable or uncertain in Med-CMH. Suppose $y_i \in \{0, 1\}$ denotes the correspondence score, which indicates that the medical pair $(\mathcal{I}_i, \mathcal{T}_i)$ is positively correlated

($y_i = 1$) or not (y_i =0). Additionally, an $N \times N$ similarity matrix $\mathcal{S}$ is generated based on labels. The main objective of our work is to obtain unified noisy-free hash codes by mapping image and text data from a high-dimensional space into a common K-bit discrete Hamming space, where K is the length of hash codes.

3.2 Cross-Modal Feature Learning

As illustrated in Fig. 2, we utilize the pre-trained VIT [6] and GPT-2 [24] as image and text encoders to extract feature representations, i.e., $\mathcal{F}_i^v = [g_i^v, Z_i^v]$, $\mathcal{F}_i^t = [g_i^t, Z_i^t]$, and $\mathcal{F}_i^c = [g_i^c, Z_i^c]$, where g represents the global class embeddings, and Z represents the sequence of local token embeddings. The obtained image captions can be considered as additional text modalities to enhance the original textual features. Moreover, we implement a weight-sharing strategy between the text and caption encoders to ensure the consistency of textual representations.

3.3 Visual-Textual Prompt Learning

By freezing the backbone network during fine-tuning, we design two prompt mechanisms, i.e., visual prompt learning (VPL) and textual prompt learning (TPL), capable of enhancing alignment and contextual awareness between the medical visual and textual representations.

Visual Prompt Learning. To learn the visual context prompts, we design a visual adapter $\mathcal{F}_v$ to transfer feature prompts from caption modality to image modality, which is implemented as a learnable linear layer to map d_c dimensional inputs to d_v. As such, the visual context prompts can be defined as $p_{i,j}^v = \mathcal{F}_v(Z_{i,j-1}^c)$. In addition, these visual context prompts are further introduced in deeper transformer layers of the image encoder alongside input image tokens,

$$[g_{i,j}^v, Z_{i,j}^v, \_] = \mathrm{Encoder_v}\big([g_{i,j-1}^v, Z_{i,j-1}^v, p_{i,j-1}^v]\big), \quad j = k, k+1, ...n, \tag{1}$$

where k represents k-th depth of transformer layers, n is the total number of layers, and "_" indicates that the output tokens are discarded.

Textual Prompt Learning. By contrast, TPL introduces a textual adapter F_t to generate textual context prompts, which consists of a down projection layer, a linear rectification function, and an up projection layer, i.e., $p_{i,j}^t = \mathcal{F}_t(Z_{i,j-1}^c)$. Then, the global text features are rectified by global caption features to enhance the noisy correspondence learning, i.e., $\overline{g}_{i,j}^t = \mathcal{F}_t(g_{i,j}^c) + g_{i,j}^t$. Moreover, the textual context prompts and the rectified global text features are incorporated in deeper transformer layers of the text encoder alongside input text tokens, i.e.,

$$[g_{i,j}^t, Z_{i,j}^t, \_] = \mathrm{Encoder_t}\big([\overline{g}_{i,j-1}^t, Z_{i,j-1}^t, p_{i,j-1}^t]\big), \quad j = k, k+1, ...n. \tag{2}$$

3.4 Noise-Robust Contrastive Learning

To further strengthen the prompt contrast of positive pairs while mitigating the impact of noisy correspondence pairs, we propose a noise-robust contrastive learning strategy that can dynamically fine-tune the intensity of contrastive learning across modalities. Specifically, we utilize a residual multi-layer perceptrons (ResMLP) to align the global embeddings $\widetilde{g}_i^{(*)}$ into the same dimensionalities, i.e., $\widetilde{g}_i^{(*)} = \mathrm{ResMLP}(g_i^{(*)})$. Based on the above preprocessing operations, the noise-robust contrastive learning strategy can be defined as follows:

$$\mathcal{L}_{\text{nrcl}} = \frac{1}{M}\sum_{i=1}^{M}\big(- \log \frac{\exp((\widetilde{g}_i^v)^\top \widetilde{g}_i^t/\widehat{\tau}_i)}{\sum_{c=1}^{M}\exp((\widetilde{g}_i^v)^\top \widetilde{g}_c^t/\widehat{\tau}_i)} - \log \frac{\exp((\widetilde{g}_i^t)^\top \widetilde{g}_i^v/\widehat{\tau}_i)}{\sum_{c=1}^{M}\exp((\widetilde{g}_i^t)^\top \widetilde{g}_c^v/\widehat{\tau}_i)}\big), \tag{3}$$

where $\widehat{\tau}_i = \tau + \gamma \cdot \mathcal{N}_i$ is the adaptive temperature parameter, τ indicates the original temperature parameter, and γ is a hyper-parameter. It is noted that $\mathcal{N}_i$ is calculated by Jensen-Shannon divergence [18] between the extracted global caption features g_i^t and global text features g_i^c,

$$\mathcal{N}_i = \mathcal{D}_{\text{JS}}^{i}(g_i^t || g_i^c) = \frac{1}{2}\big(\sum_j g_{ij}^t \log \frac{2g_{ij}^t}{g_{ij}^t + g_{ij}^c} + \sum_j g_{ij}^c \log \frac{2g_{ij}^c}{g_{ij}^t + g_{ij}^c}\big) \in [0,1]. \tag{4}$$

3.5 Cross-Modal Hashing Learning

The purpose of Med-CMH is to map features into the Hamming space, ensuring that the distance relationships between hash codes reflect the semantic similarity of different modalities. To this end, we adopt a localized token aggregation (LTA) strategy with a two-layer transformer [22] to localize the preservation of the most crucial implicit semantic knowledge from the global embeddings to form the local embeddings. Then, a hashing linear projection layer with the *tanh* activation function is designed to map local embeddings to local semantic features $f_i^{(*)}$ in the K-bit Hamming space. Similarly, we introduce a hashing linear projection layer to decompose the projected features into global semantic features $h_i^{(*)}$. Both global and local features are integrated to learn the unified hash code,

$$b_i = \text{sign}(\lambda(h_i^v + h_i^t) + (1-\lambda)(f_i^v + f_i^t)), \quad \lambda \in [0,1], \tag{5}$$

where λ denotes a learnable hyper-parameter. Moreover, the quantization loss is used to learn uniform compact hash codes,

$$\mathcal{L}_{\text{quan}} = \frac{1}{KM}\sum_{i=1}^{M}\big(\|b_i - \frac{1}{2}(h_i^v + f_i^v)\|_2^2 + \|b_i - \frac{1}{2}(h_i^t + f_i^t)\|_2^2\big), \tag{6}$$

where M is the batch size. In addition, the intra-modal similarity preservation loss aims to preserve semantic similarities within modalities,

$$\mathcal{L}_{\text{intra}} = -\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\varOmega_{ij}^{(*)} - \log\left(1 + e^{\varOmega_{ij}^{(*)}}\right)\right), \tag{7}$$

where $\varOmega_{ij}^{(*)} = \frac{1}{2}(f_i^{(*)})^T f_j^{(*)}$ indicates the inner product among local semantic representations. Moreover, inter-modal similarity preservation is designed to preserve semantic similarities across modalities, that is,

$$\mathcal{L}_{\text{inter}} = -\frac{1}{MN}\Big(\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\varTheta_{ij} - \log\left(1 + e^{\varTheta_{ij}}\right)\right) + \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathcal{S}_{ij}\varPhi_{ij} - \log\left(1 + e^{\varPhi_{ij}}\right)\right)\Big), \tag{8}$$

where $\varTheta_{ij} = \frac{1}{2}(h_i^t)^T h_j^v$ and $\varPhi_{ij} = \frac{1}{2}(h_i^v)^T h_j^t$ denotes the inner product among global semantic representations. Overall, the training objective of our MCPH approach encompasses a combination of various loss functions, i.e.,

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{quan}} + \beta \cdot \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{nrcl}}, \tag{9}$$

where α and β are trade-off hyper-parameters during the training phase.

Table 1. Comparisons of mAP (%) scores on Open-I with different noise rates (NR). The best and second-best results are highlighted in bold and underlined.

	Methods	Ref.	NR:0%			NR:20%			NR:50%		
			16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I → T	DCHMT	MM'22	41.04	42.22	42.25	38.64	39.38	40.01	35.37	35.69	35.48
	DAPH	SIGIR'23	54.45	54.47	53.26	47.36	50.28	52.41	45.38	43.01	45.22
	MITH	MM'23	60.15	58.96	58.95	53.22	52.17	53.98	44.79	44.25	47.43
	MCPH	Ours	**62.86**	**63.42**	**64.04**	**58.24**	**59.13**	**60.44**	**54.24**	**54.66**	**55.94**
		↑	+2.71	+4.46	+5.09	+5.02	+6.96	+6.46	+8.86	+10.41	+8.51
T → I	DCHMT	MM'22	46.15	46.81	47.49	43.29	43.90	44.35	37.65	36.34	36.14
	DAPH	SIGIR'23	56.57	58.99	59.28	51.38	55.71	56.85	44.50	49.03	51.56
	MITH	MM'23	63.66	62.43	62.97	59.31	59.29	57.72	53.73	53.50	52.54
	MCPH	Ours	**66.13**	**66.11**	**66.73**	**62.41**	**63.37**	**62.47**	**58.56**	**58.50**	**58.99**
		↑	+2.47	+3.68	+3.76	+3.10	+4.08	+4.75	+4.83	+5.00	+6.45

Table 2. Comparisons of mAP (%) scores on MIMIC-CXR with different noise rates (NR). The best and second-best results are highlighted in bold and underlined.

	Methods	Ref.	NR:0%			NR:20%			NR:50%		
			16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I → T	DCHMT	MM'22	56.52	59.63	59.76	52.43	55.63	55.77	50.62	53.62	53.70
	DAPH	SIGIR'23	60.32	62.00	62.11	58.51	60.56	61.51	55.25	55.76	56.89
	MITH	MM'23	59.27	60.55	61.56	60.40	61.80	58.26	57.51	59.91	58.32
	MCPH	Ours	**66.21**	**66.03**	**66.64**	**65.07**	**65.51**	**65.74**	**64.62**	**64.71**	**64.99**
		↑	+5.89	+4.03	+4.53	+4.67	+3.71	+4.23	+7.11	+4.80	+6.67
T → I	DCHMT	MM'22	55.74	59.46	59.45	51.85	54.69	55.02	50.07	52.81	52.89
	DAPH	SIGIR'23	60.02	61.40	61.85	57.12	58.83	59.69	55.39	56.34	56.91
	MITH	MM'23	60.13	60.20	58.09	58.00	59.40	60.17	56.34	58.68	58.87
	MCPH	Ours	**63.89**	**64.35**	**64.94**	**63.03**	**63.50**	**63.81**	**62.55**	**62.73**	**64.04**
		↑	+3.76	+2.95	+3.09	+5.03	+4.10	+3.64	+6.21	+4.05	+5.17

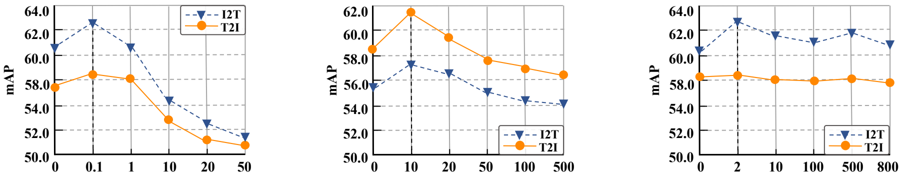
4 Experiments

4.1 Datasets & Baselines & Evaluation Metrics

To comprehensively evaluate the performance of our method, we conduct comparative results on two widely-used benchmark datasets, i.e., Open-I and MIMIC-CXR. In our experiments, we randomly select 500 pairs in Open-I [5] as the query set, the remaining 2,318 pairs as the retrieval set, and 1,000 pairs sampled from the retrieval set to form the training set. From the MIMIC-CXR [11] dataset, 2,000 pairs are randomly sampled as the query set, the rest 87,286 pairs formed the retrieval set, and 10,000 pairs for training. We compare the proposed MCPH method with several state-of-the-art transformer-based methods, including DCHMT [28], DAPH [29], and MITH [22]. Following [22], we adopt the mean average precision (mAP) as the comparison metric for two cross-modal retrieval tasks, i.e., image-to-text retrieval (I → T) and text-to-image retrieval (T → I). By randomly shuffling the training images and texts, we compare various methods under different noise rates, i.e., 20% and 50%.

Table 3. Ablation studies of mAP (%) on Open-I using 16 bits with 20% noise rate, where Variant-I indicates the BASE model, Variant-II denotes "BASE+TPL", Variant-III represents "BASE+VPL", and Variant-IV indicates "BASE+VPL+TPL".

Variants	Variant-I	Variant-II	Variant-III	Variant-IV	MCPH
I → T	54.56	55.45	56.74	57.57	**58.24**
T → I	56.24	60.34	60.77	61.39	**62.41**



(a) Parameter α (b) Parameter β (c) Parameter $\gamma/100$

Fig. 3. Parameter analysis (%) on the Open-I dataset with 20% noise rate.

4.2 Comparisons with State-of-The-Arts

In this part, we evaluate the performance of the proposed MCPH framework on the Open-I and MIMIC-CXR datasets with mAP scores under different noise rates, and the comparative results are summarized in Table 1 and Table 2. From these tables, we can observe that the proposed MCPH significantly outperforms all state-of-the-art baselines with all cases on two datasets. Specifically, our MCPH achieves the highest mAP scores on Open-I and MIMIC-CXR datasets with all hash code lengths in conventional settings, which can verify the effectiveness of our MCPH for addressing Med-CMH tasks. Due to the influence of noisy correspondence, almost all comparative methods suffer from varying degrees of performance degradation under different noise rates. Nonetheless, our MCPH not only achieves the best overall performance under low noise but also exhibits superior robustness under relatively high noise. When the noise rate is 50%, the proposed MCPH framework surpasses the second-best method on MIMIC-CXR by the mean mAP of 4.82%, 4.20%, 6.19% for I → T, and 3.27%, 4.26%, 5.14% for T → I, respectively. The above comparative results consistently demonstrate the superiority and generalizability of our MCPH framework.

4.3 Ablation Studies & Parameters Analysis

We conduct comprehensive ablation studies by systematically evaluating the impact of each component in MCPH. As shown in Table 3, we can conclude that all designed modules can complement and reinforce each other, which further verifies the combined effects of our MCPH method. Moreover, we empirically perform parameter analyses to analyze the effects of three key hyper-parameters, i.e., α, β, and γ, as shown in Eq. (9) and Eq. (3). The comparative results are illustrated in Fig. 3. It can be observed that our MCPH method achieves the best performance when $\alpha = 0.1$, $\beta = 10$, and $\gamma = 200$, respectively. Hence, we can summarize that our MCPH can obtain preferable performance by an optimal combination of these hyperparameters. More experimental results and implementation details are shown in the supplementary material.

5 Conclusion

This paper identified a widely-exist but rarely-explored problem in Med-CMH, i.e., noisy correspondence. To overcome this challenge, we presented a robust and reliable medical cross-modal prompt hashing that benefits greatly from the potential synergistic efficacy of multi-modal prompt optimization with noise-robust contrastive constraint for embracing the robustness against noisy correspondence. Furthermore, extensive experiments were carried out to validate the effectiveness of our MCPH framework in mitigating noisy correspondence.

Acknowledgments. This work was supported in part by the NSFC fund (Grant NOs. 62176077, 62302172), in part by the Shenzhen Key Technical Project (Grant NOs. JSGG20220831092805009, JSGG20220831105603006, JSGG20201103153802006, KJZD20230923115117033), in part by the Guangdong International Science and Technology Cooperation Project (Grant NO. 2023A0505050108), in part by the Shenzhen Fundamental Research Fund (Grant NO. JCYJ20210324132210025), and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant NO. 2022B1212010005).

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

References

1. Bai, C., Zeng, C., Ma, Q., Zhang, J., Chen, S.: Deep adversarial discrete hashing for cross-modal retrieval. In: Proceedings of ICMR. pp. 525–531 (2020)
2. Chen, B., Ye, Z., Liu, Y., Zhang, Z., Pan, J., Zeng, B., Lu, G.: Combating medical label noise via robust semi-supervised contrastive learning. In: Proceedings of MICCAI. pp. 562–572 (2023)
3. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of AAAI. pp. 1–10 (2024)
4. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024)
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of ICLR (2021)
7. Han, H., Miao, K., Zheng, Q., Luo, M.: Noisy correspondence learning with meta similarity correction. In: Proceedings of CVPR. pp. 7517–7526 (2023)
8. Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C.: Remind your neural network to prevent catastrophic forgetting. In: Proceedings of ECCV. pp. 466–483 (2020)
9. Huang, Z., Niu, G., Liu, X., Ding, W., Xiao, X., Wu, H., Peng, X.: Learning with noisy correspondence for cross-modal matching. In: Proceedings of NeurIPS. vol. 34, pp. 29406–29419 (2021)
10. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: Proceedings of CVPR. pp. 3232–3240 (2017)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
12. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of CVPR. pp. 19113–19122 (2023)
13. Kumar, A., Kim, J., Wen, L., Fulham, M., Feng, D.: A graph-based approach for the retrieval of multi-modality medical images. Medical Image Analysis (MIA) **18**(2), 330–342 (2014)
14. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of EMNLP. pp. 3045–3059 (2021)
15. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of CVPR. pp. 4242–4251 (2018)
16. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of IJCNLP. pp. 4582–4597 (2021)
17. Li, Z., Zhang, X., Müller, H., Zhang, S.: Large-scale retrieval for medical image analytics: A comprehensive review. Medical Image Analysis (MIA) **43**, 66–84 (2018)

18. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory (TIT) **37**(1), 145–151 (1991)
19. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: Proceedings of CVPR. pp. 3864–3872 (2015)
20. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. In: Proceedings of ACL. pp. 1–11 (2021)
21. Liu, X., Tang, W., Lu, J., Zhao, R., Guo, Z., Tan, F.: Deeply coupled cross-modal prompt learning. In: Proceedings of ACL. pp. 7957–7970 (2023)
22. Liu, Y., Wu, Q., Zhang, Z., Zhang, J., Lu, G.: Multi-granularity interactive transformer hashing for cross-modal retrieval. In: Proceedings of ACM MM. pp. 893–902 (2023)
23. Qin, Y., Peng, D., Peng, X., Wang, X., Hu, P.: Deep evidential learning with noisy correspondence for cross-modal retrieval. In: Proceedings of ACM MM. pp. 4948–4956 (2022)
24. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
25. Shen, H.T., Liu, L., Yang, Y., Xu, X., Huang, Z., Shen, F., Hong, R.: Exploiting subspace relation in semantic labels for cross-modal hashing. IEEE Transactions on Knowledge and Data Engineering (TKDE) **33**(10), 3351–3365 (2020)
26. Shi, Z., Lipani, A.: Don't stop pretraining? make prompt-based fine-tuning powerful learner. In: Proceedings of NeurIPS. vol. 36, pp. 1–18 (2024)
27. Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M.: Interpretability-guided content-based medical image retrieval. In: Proceedings of MICCAI. pp. 305–314 (2020)
28. Tu, J., Liu, X., Lin, Z., Hong, R., Wang, M.: Differentiable cross-modal hashing via multimodal transformers. In: Proceedings of ACM MM. pp. 453–461 (2022)
29. Tu, R.C., Mao, X.L., Ji, W., Wei, W., Huang, H.: Data-aware proxy hashing for cross-modal retrieval. In: Proceedings of SIGIR. pp. 686–696 (2023)
30. Xu, L., Zeng, X., Zheng, B., Li, W.: Multi-manifold deep discriminative cross-modal hashing for medical image retrieval. IEEE Transactions on Image Processing (TIP) **31**, 3371–3385 (2022)
31. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.: Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Transactions on Image Processing (TIP) **26**(5), 2494–2507 (2017)
32. Yang, E., Liu, M., Yao, D., Cao, B., Lian, C., Yap, P.T., Shen, D.: Deep bayesian hashing with center prior for multi-modal neuroimage retrieval. IEEE Transactions on Medical Imaging (TMI) **40**(2), 503–513 (2020)
33. Yang, S., Xu, Z., Wang, K., You, Y., Yao, H., Liu, T., Xu, M.: Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In: Proceedings of CVPR. pp. 19883–19892 (2023)
34. Zhang, D., Li, W.J.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of AAAI. vol. 28, pp. 1–7 (2014)
35. Zhang, Y., Ou, W., Shi, Y., Deng, J., You, X., Wang, A.: Deep medical cross-modal attention hashing. World Wide Web **25**(4), 1519–1536 (2022)
36. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of CVPR. pp. 16816–16825 (2022)
37. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) **130**(9), 2337–2348 (2022)