Advancing In-Context Learning for Efficient and Stable Medical Report Generation

Mingjie Li, Rui Liu, Zeyi Shi, Mingfei Han, Lina Yao, *Senior Member, IEEE*, Zhihui Li, Xiaojun Chang, *Senior Member, IEEE*, Kilian M. Pohl, Md Tauhidul Islam, and Lei Xing

***Abstract*—Vision-language models (VLMs) have shown strong generalization across multimodal tasks, but adapting them to medical report generation (MRG) often demands extensive paired image-text data that are limited due to data privacy and annotation cost. In-context learning (ICL) offers a promising training-free alternative, yet standard ICL approaches rely on long demonstration prompts that are computationally inefficient and often yield inconsistent or clinically inaccurate descriptions. To address these challenges, we propose Principal In-Context Vectors (PCVs), a compact latent-guidance framework that distills multimodal demonstrations into stable semantic representations. By extracting hidden states from auto-regressive VLMs and applying principal component analysis (PCA), we identify robust semantic directions that remain stable under input perturbations. These PCVs are then injected into new queries to steer generation toward accurate and clinically meaningful outputs without any model tuning. Extensive experiments on four MRG benchmark datasets show that our approach can enhance both zero-shot and fully supervised generation quality across diverse settings, including cross-center, cross-disease, and longitudinal scenarios. This work provides a lightweight and scalable approach to adapt pre-trained VLMs for practical clinical deployment.**

Received 2 June 2025; revised 10 March 2026; accepted 26 April 2026. Date of publication 4 May 2026; date of current version 6 August 2026. This work was partly supported by the National Institute of Health under Grant DA057567 and Grant AA021697, in part by the Stanford HAI Google Cloud Credits, and in part by the 2024 Stanford HAI Hoffman-Yee Grant. Recommended for acceptance by S. Wang. *(Mingjie Li and Rui Liu are contributed equally to this work.) (Corresponding authors: Lei Xing; Kilian M. Pohl; Zhihui Li.)*

Mingjie Li was with the Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305 USA. He is now with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: mingjie.li@sjtu.edu.cn).

Rui Liu and Zeyi Shi are with the Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: rui.liu-8@student.uts.edu.au; zeyi.shi@student.uts.edu.au).

Mingfei Han, Zhihui Li, and Xiaojun Chang are with the University of Science and Technology of China, Hefei 230052, China (e-mail: hmf282@gmail.com; lizhihuics@ustc.edu.cn; xjchang@ustc.edu.cn).

Lina Yao is with the CSIRO's Data6, Sydney, NSW 2015, Australia, and also with the University of New South Wales, Sydney 2052, Australia (e-mail: lina.yao@data61.csiro.au).

Kilian M. Pohl is with the Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305 USA, and also with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: kpohl@stanford.edu).

Md Tauhidul Islam and Lei Xing are with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: tauhid@stanford.edu; lei@stanford.edu).

The code is available at https://github.com/rliu0016/PCVs.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2026.3689780, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2026.3689780

***Index Terms*—In-context learning, principal component analysis, vision language model, and medical report generation.**

I. Introduction

MEDICAL report generation (MRG) aims to automatically generate free-text diagnostic reports from images in radiology, cardiology, ophthalmology [1], [2]. It serves as a critical step in building reliable, AI-assisted clinical decision-making systems, potentially reducing the documentation burden on clinicians while enhancing reporting efficiency and consistency [3]. Recent advances in large vision-language models (VLMs), such as GPT-4o [4], LLaVA [5], and Gemini [6], have demonstrated remarkable capabilities in multimodal reasoning and generation, sparking growing interest in applying VLMs to medical domains [3], [7]. Consequently, a growing number of state-of-the-art (SOTA) MRG methods have adopted VLM-based architectures, leveraging various types of visual prompts to guide large language models (LLMs) in producing clinically meaningful reports. For instance, RGRG [8] and MATM [9] utilizes object detectors to extract regional features, R2GPT [10] and CheXpert Plus [11] employ Swin [12] backbones to encode patch-level embeddings, and PromptMRG [13] and CoFE [14] incorporate structured disease labels directly into the prompt design. These approaches have yielded promising results compared to traditional task-specific models [15], [16], confirming the potential of vision-language reasoning in radiology.

However, existing methods rely heavily on large-scale supervised training or task-specific fine-tuning, which incurs substantial computational costs and limits scalability. As increasingly heterogeneous and high-quality MRG datasets become publicly available [11], [17], [18], [19], we anticipate a new wave of general-purpose and publicly available VLMs tailored for clinical imaging tasks. In this context, a critical question arises: ***how can we efficiently and reliably adapt such models toward specific downstream tasks without additional training?*** This need for efficient adaptation becomes even more pressing given the variability inherent in real-world medical reporting scenarios. In practice, differences in institutional writing styles, reporting protocols, and disease distribution across clinical centers introduce significant domain shifts [20]. In addition, downstream tasks such as the generation of longitudinal reports, further complicate the report generation process as they require models to reason about patient history and temporal changes [21], [22]. Such

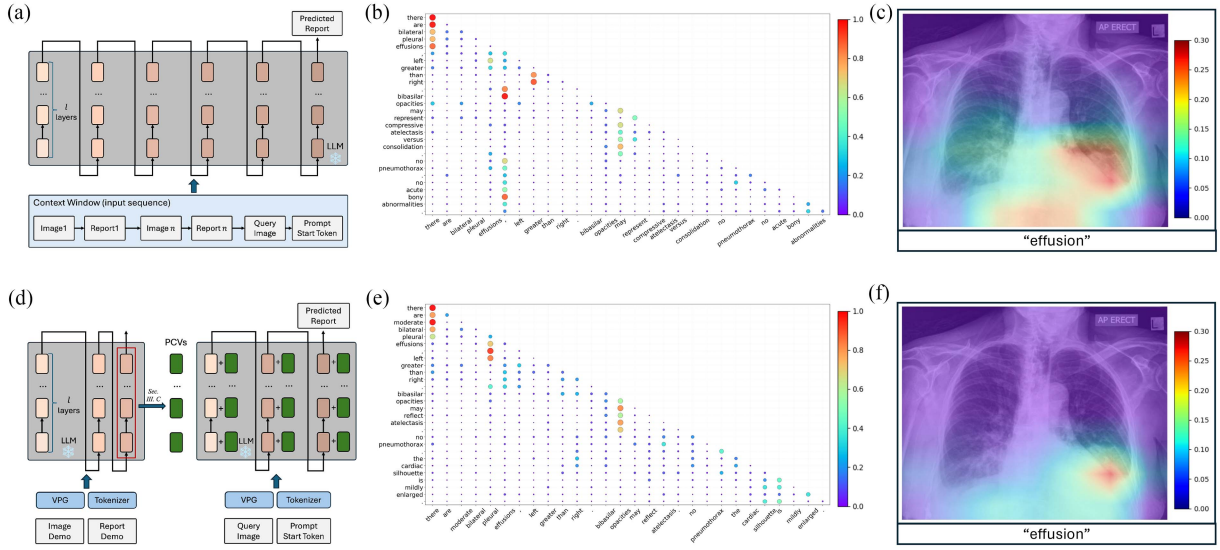


Fig. 1. Traditional in-context learning (a) injects the context window consisting of several demonstrations, a query image, and prompt start tokens, but attention often focuses on non-semantic tokens (b: self-attention matrices for each generated token) and diffuse image regions (c: cross-attention weights on the input X-ray image, highlighting the token "*effusion*"). Our PCV paradigm (d) steers the model through a semantic direction in latent space, yielding concentrated attention on key medical terminology (e) and localized image regions (f). VPG refers to the visual prompt generator.

heterogeneity challenges the generalizability of existing VLMs and highlight the importance of training-free, task-adaptive mechanisms that can robustly guide generation without requiring additional supervision.

One promising solution to this challenge is in-context learning (ICL) [23], [24], which enables a model to adapt to a new task or domain at inference time by conditioning on a small set of input–output demonstrations [25]. However, applying ICL to vision-language models in the medical domain presents unique challenges. First, ICL inefficiently handle extremely long medical text inputs such as Radiology reports [26], which are often detailed and verbose so that they require hundreds of tokens for each image-report demonstration. Concatenating multiple such demonstrations into the input sequence leads to substantial memory and computation overhead during inference [27]. Second, the performance of ICL in MRG is often highly sensitive to the choice and formatting of demonstrations [27]. The quality and consistency of generated reports can vary significantly depending on the structure, ordering, and content of the few-shot examples. This sensitivity is further amplified in multi-modal contexts, where mismatches between visual and textual semantics can degrade alignment [28]. Third, hallucinations in VLMs often stem from the instability caused by minor perturbations to visual representations, which disrupts consistent vision–language alignment during decoding [29], [30]. The use of ICL can further exacerbate hallucinations by introducing additional visual–textual inputs that amplify representational noise. Even slight variations, *e.g.* changes in masking or example ordering, can shift the embedding space and increase the risk of semantic drift, particularly in long-form generation tasks. For example, Fig. 1(b) shows the self-attention distribution of CheXpert Plus benchmark model (CXP) during report generation. In the later decoding steps, the model's attention becomes disproportionately focused on the comma token, rather than on clinically meaningful terms. This phenomenon reflects a breakdown in contextual coherence and suggests that the model loses semantic grounding as the generation progresses, leading to factual hallucinations.

To address the challenges of inefficiency and sensitivity, we previously proposed a compact and training-free alternative based on Multimodal Context Vectors (MCVs) [20]. The core idea is to leverage the auto-regressive nature of LLMs: rather than feeding full image–report pairs into the prompt Fig. 1(a), we extract the final hidden state of each demonstration, which implicitly encodes its multimodal contextual semantics. These MCVs then serve as compressed representations of individual demonstrations. During inference, the averaged MCVs from all pairs is injected into the latent space of the target input, enabling the VLM to incorporate contextual knowledge and generate downstream outputs in a more efficient manner. Building upon this foundation, this work introduces Principal Context Vectors (PCVs) to further tackle the hallucination problem and enhance the stability of medical report generation Fig. 1(d).

To construct PCVs, we introduce controlled perturbations (i.e., random masking) to the input image to simulate diverse but semantically equivalent views. For each perturbed image, we extract the MCVs from the final hidden state of the VLM's decoder, capturing the latent alignment between the visual and textual modalities. While each individual MCV may vary due to non-semantic input noise, collectively they reflect the underlying semantic intent of the demonstration. This design is motivated by the observation that hallucinations in VLMs often arise from the instability of visual-language representations under minor perturbations [27], [31], [32], [33]. Even slight shifts in the visual input can misalign the latent space and cause the model to rely on syntactically trivial or clinically irrelevant cues during generation. To mitigate this, we apply Principal Component Analysis (PCA) [34] across the set of MCVs to extract the

dominant directions of semantic consistency. The top principal components, which capture stable, cross-perturbation context signals, are aggregated to form the PCV. This PCV is then injected into the query sample's latent space to steer the VLMs. Compared to just using MCVs, doing so results in the VLMs to generate more accurate and semantically consistent reports (Fig. 1(b) and (e)) as the model now focuses on clinically salient terminology instead of punctuation tokens. Moreover, cross-attention visualizations (Fig. 1(c) and (f)) show that PCVs reduce attention dispersion over irrelevant image regions and encourage more focused alignment with clinically meaningful areas.

To comprehensively evaluate the effectiveness of PCVs, we perform three experiments targeting key challenges in MRG. First, we assess the ability of PCVs to mitigate hallucinations by analyzing attention behavior and factual correctness in generated reports under controlled perturbations. Second, we evaluate the zero-shot generalization performance of PCV-guided VLMs across both cross-center and cross-disease scenarios, highlighting the adaptability of the framework to heterogeneous clinical environments without fine-tuning. Third, we test the performance of PCVs-guided models in longitudinal report generation, a more complex task that requires temporal reasoning and consistency over time. Experiments conducted on three publicly available benchmarks, IU X-Ray [35], MIMIC-CXR [17], CheXpert Plus [11] and Peir Gross [36], demonstrate that our method enables VLMs to achieve competitive medical report generation performance without retraining, using only a few in-context examples. Our main contributions are as follows:

- We propose PCVs, a training-free mechanism that injects stable semantic guidance into VLMs by applying PCA over perturbed in-context representations.
- We provide a novel strategy to reduce hallucinations in MRG by enhancing visual-language alignment at the latent level.
- We conduct extensive evaluations on two real-world zero-shot settings and longitudinal reporting across five datasets.

Compared to our earlier conference version, this journal paper substantially broadens both the methodological scope and the empirical validation. While the conference paper first introduced MCVs [20] for zero-shot MRG, the present work advances this framework through a principled formulation of PCVs. PCVs extract perturbation-invariant semantic directions from MCVs via PCA, providing a theoretically grounded and training-free mechanism that enhances generation stability and mitigates hallucinations in VLMs. We further include a formal derivation demonstrating that PCA yields a provably optimal, robustness-enhancing projection under perturbations. Beyond methodological refinement, we expand the empirical coverage by incorporating two new benchmark models, two new datasets, and a new experimental setting to evaluate temporal consistency in follow-up studies. Moreover, PCVs provide statistically significant improvements over standard MCVs in both report quality and robustness metrics. These contributions collectively enable more reliable and adaptable zero-shot MRG across diverse medical settings.

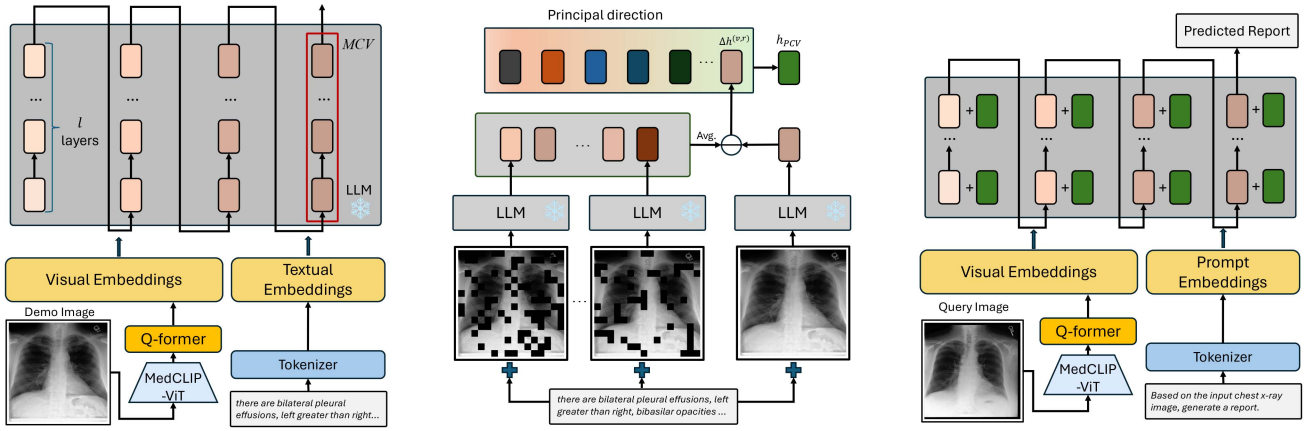
II. Related Work

A. Vision Language Models in Medical Report Generation

Most MRG models adopt an encoder–decoder architecture, where a visual encoder extracts image features and a language decoder generates free-text reports based on these representations [37]. The evolution of MRG methods has been closely tied to advancements in language decoders, largely due to the unique complexity of medical reports. Unlike general image captioning, which typically involves a single sentence describing visual content, medical reports are long-form documents comprising multiple sentences, each covering distinct clinical observations or anatomical regions. Early MRG models employed dual-layer LSTM-based decoders [15], [38], which were later replaced by transformer-based architectures [39], [40], [41] for improved contextual modeling. Subsequent works introduced contrastive learning [42], [43], [44], [45] to pretrain the Transformer-based decoder, aiming to better align visual and textual modalities before report generation. More recently, large-scale pretrained VLMs have been introduced in medical reporting tasks, inspired by their success across a wide range of multimodal generation tasks in the natural domain, such as visual question answering, image captioning, vision-language navigation, and document understanding [46], [47].

Recent MRG approaches leveraging VLMs can be broadly categorized into three types based on how they incorporate visual features. The first one adopts the classical VLM architecture, where a vision transformer serves as the image encoder and a large language model functions as the decoder. For example, R2GPT [10] employs a Swin Transformer [12] to extract patch-level visual features and pairs it with a frozen LLaMA-3 decoder for report generation. They also explored lightweight fine-tuning strategies such as LoRA [25] to further enhance generation quality. Similarly, CXP [11] utilizes a Swin-based vision encoder but opts to train a custom BERT-based decoder [48], enabling tighter integration with domain-specific language distributions. The second one focuses on leveraging regional features to capture finer-grained visual semantics and enable localized report generation. For instance, RGRG [8] first trains an object detector to localize candidate regions within the image and classify each as normal or abnormal. For abnormal regions, a dedicated GPT-2 [49] decoder is employed to generate corresponding textual descriptions. Similarly, Chen et al. [1] utilize spatial masks derived from a universal segmentation module to isolate and encode local image features associated with specific referring expressions. The third line of research leverages visual features to generate disease-related prompts, which are then fed into LLMs to guide report generation. For example, Jin et al. [13] first predict disease-related labels that are then formatted into structured textual prompts to steer the LLM's generation process. Along the same line, Li et al. [14] design prompts with counterfactual explanations, aiming to teach the model not only what is present in the image but also what is absent.

While these VLM-based approaches have achieved promising results, they typically require substantial computational resources for training or fine-tuning, especially when scaling



(a) Extract MCV by our R2VLM. (b) Calculate PCV. (c) Apply PCV.

Fig. 2. Steering R2VLM with our proposed PCV consists of three key stages: (a) For each multimodal demonstration, we perform a forward pass through the R2VLM to obtain MCVs, i.e., all hidden states from each layer at the last token. (b) We apply controlled perturbations to the same image, generate multiple MCVs, and apply principal component analysis to identify the dominant semantic directions, yielding the PCV. (c) The PCV is injected into the latent space of new query samples to guide report generation.

to large models and datasets. As high-quality medical report datasets become increasingly available, we anticipate that VLM-based pipelines will evolve toward modular and reusable architectures, where pretrained visual and language components are combined flexibly across tasks. In this context, developing training-free and computation-efficient methods for effectively leveraging VLMs becomes more important, offering a practical and scalable direction for real-world clinical deployment.

B. *In-Context Learning (ICL)*

ICL has emerged as a powerful paradigm for adapting LLMs to new tasks without parameter updates by conditioning the model on a small set of demonstrations provided within the same input sequence as the query [23]. This approach enables efficient and flexible task transfer as the model infers intent and structure from contextual examples rather than relying on explicit fine-tuning. However, the effectiveness of ICL is highly sensitive to the choice, formatting, and ordering of demonstrations—particularly in multimodal settings [28], where the alignment between visual and textual content further complicates contextual inference. To better understand and manipulate the underlying mechanisms of ICL, recent works have proposed the use of compact vector representations, such as task vectors [50], function vectors [51], and in-context vectors [27], that summarize the influence of demonstrations within the model's latent space. For instance, Hendel et al. [50] show that a training set can be distilled into a single task vector, while Todd et al. [51] reveal that LLMs encode functional mappings that are compositional and robust across contexts. Building on these insights, our prior work introduced MCVs, which encode the contextual semantics of each image–text demonstration by extracting the final hidden state from the VLM's decoder. In this paper, we further advance this line of research by introducing the concept of PCVs, which identifies semantically stable directions across perturbed demonstrations. To the best of our knowledge, this is the first attempt to leverage principal latent representations for improving training-free multimodal adaptation in medical report generation.

III. Methodology

The proposed pipeline for extracting and applying PCVs to guide VLMs for efficient and stable medical report generation is illustrated in Fig. 2. The goal is to leverage the latent semantics encoded in multimodal demonstrations to steer the generation process without retraining. In this section, we first describe the backbone architecture and the procedure for extracting MCVs (Fig. 2(a)), which was inspired by our findings. We then introduce the Principal In-Context Vector Construction (PCV, Fig. 2(b)), which provides a theoretical formulation for deriving robust semantic directions via PCA. In the end, we explain PCV-guided generation mechanism (Fig. 2(c)), which details how these vectors are injected into query samples to improve stability and factual consistency during report generation.

A. *Backbone*

A task-specific VLM backbone (a.k.a, **R2VLM**, Fig. 2(a)) is constructed to support report generation in a modular and adaptable manner. Compared to prior Med-VLMs that typically use shallow MLP-based visual prompt generators (VPGs), our design incorporates a Q-Former [52], [53] as the VPG. This enables better alignment between image and language modalities than a simple MLP and also allows us to assess the robustness of our PCV method under a more expressive visual encoding pipeline.

Specifically, we adopt MedCLIP-ViT [54] as the vision encoder to extract image features, denoted as Img. These features are then passed to the Q-Former, which transforms them into a set of visual tokens aligned in dimensionality with the LLM's text embeddings. To reduce input redundancy and improve efficiency, we retain a small number of learned visual queries.

For the language model, we integrate LLaMA-2-7B-chat [55]. Visual and textual inputs are concatenated and passed directly to the LLM without any architectural modification, following the simple token-level fusion strategy used in Flamingo [56]. A special token `[IMG]` is inserted to help the LLM distinguish visual context from natural language input.

Pre-training Paradigm: We pre-train our Med-LMM using an instruction-tuning setup that preserves the original auto-regressive objective of the LLM. Given a report X_r of length L_r, an instructional prompt X_p, and visual embeddings X_v, the model is optimized using the negative log-likelihood over the report tokens:

$$\mathcal{L}(\theta; X_p, X_v, X_r) = -\sum_{i=1}^{L_r} \log p_\theta \left(X_i \mid X_p, X_v, X_{r<i}\right), \quad (1)$$

where θ denotes the trainable parameters and $X_{r<i}$ represents the report prefix up to position $i-1$. The prompt X_p is fixed as: *"Generate a comprehensive and precise report for this Chest X-ray image."* During training, the LLM is frozen, and only parameters in the VPG are updated.

B. Multimodal Context Vector Extraction

Since medical reports are substantially longer than natural image captions (often 450–700 tokens), concatenating multiple demonstrations into a single input quickly becomes infeasible. Given a pre-trained VLM with report generation capability, we extract MCVs from a set of image–text demonstrations. Leveraging the auto-regressive structure of the LLM, we obtain hidden states that inherently capture the semantic dependencies among the input prompt, visual features, and report tokens.

Given $X = [X_p, X_v, X_r]$, we extract the final hidden state at the last report token from each of the LLM's L layers as the MCV $h_{\text{MCV}} = \{h_1, h_2, \ldots, h_L\}, \quad h_i \in \mathbb{R}^d$. In our setting, the LLM consists of $L = 32$ layers and its hidden size is $d = 4096$.

C. Principal In-Context Vector Construction

PCVs (Fig. 2(b)) are derived through principal component analysis (PCA) over perturbation-induced latent variations to capture stable, perturbation-invariant semantic directions. We first formalize how perturbation-based latent shifts are computed from MCVs and then show how PCA identifies the dominant direction of semantic consistency. Then, we present a theoretical justification demonstrating that this projection provides the optimal low-rank approximation that maximizes semantic signal-to-noise ratio (SNR) and bounds information loss in the latent space.

Given a demonstration consisting of an image-text pair (v, r), we compute the perturbation-based latent shift by first obtaining its original MCVs $h^{(v,r)}$ by passing the raw input through the R2VLM:

$$h_l^{(v,r)} := f_l\left(X_{\text{clean}}^{(v,r)}\right), \quad (2)$$

where $f(\cdot)$ denote the R2VLM forward function and $f_l(\cdot)_t$ returns the d-dimensional hidden state at layer l. We then apply random masking to generate P perturbed versions of the same image-text pair and compute the average MCVs $f_l(X_{\text{masked}}^{(v,r,p)})$ under these perturbations:

$$\bar{h}_l^{(v,r)} := \frac{1}{P}\sum_{p=1}^{P} f_l\left(X_{\text{masked}}^{(v,r,p)}\right). \quad (3)$$

The latent shift vector between the original and perturbation-averaged MCVs is defined as:

$$\Delta h_l^{(v,r)} := \bar{h}_l^{(v,r)} - h_l^{(v,r)}. \quad (4)$$

This vector reflects the direction in latent space that adjusts the original MCV toward a more stable, perturbation-invariant representation. For a batch of N demonstrations, we collect all latent shift vectors at a given layer l into a matrix:

$$M_l = \begin{bmatrix} \Delta h_l^{(1)} \\ \Delta h_l^{(2)} \\ \vdots \\ \Delta h_l^{(N)} \end{bmatrix} \in \mathbb{R}^{N\times d}. \quad (5)$$

We then compute the empirical covariance of perturbation-induced shifts:

$$\Sigma = \frac{1}{N} M_l^\top M_l \in \mathbb{R}^{d\times d}. \quad (6)$$

Let $\Sigma = V\Lambda V^\top$, where $\Lambda = \text{diag}(\lambda_1 \geq \cdots \geq \lambda_d)$. The *Principal Consistent Vector* h_{PCV} is defined strictly as the single dominant eigenvector:

$$h_{\text{PCV}} = \arg\max_{\|\mu\|_2=1} \mu^\top \Sigma \mu = v_1. \quad (7)$$

According to the Eckart–Young–Mirsky theorem, PCA provides the unique rank-1 projection minimizing the reconstruction error:

$$\min_{\text{rank}(\Pi)=1} \|M_l - M_l\Pi\|_F^2 = \sum_{i=2}^{d} \lambda_i. \quad (8)$$

Hence, the variance retained by this dominant principal component is $R_1 = \frac{\lambda_1}{\sum_{i=1}^{d}\lambda_i}$, and the discarded variance (information loss) is $L_1 = 1 - R_1$. Empirically, we observe $R_1 > 0.93$ across layers, indicating that the single vector v_1 preserves over 93% of the total perturbation energy while filtering out high-frequency noise components, making higher-order aggregations unnecessary. Consider the perturbation decomposition $\Delta h = s\mu + \varepsilon$, where μ is the semantic shift direction and ε denotes zero-mean nuisance noise. The covariance of perturbations is

$$\Sigma = s^2\mu\mu^\top + \Sigma_\varepsilon. \quad (9)$$

Under the classical spiked covariance model, the leading eigenvector v_1 aligns with μ up to an angular deviation bounded by the Davis–Kahan theorem:

$$\sin\angle(v_1, \mu) \leq \frac{\|\Sigma_\varepsilon\|_2}{s^2 - \lambda_2(\Sigma)}. \quad (10)$$

When semantic variance s^2 dominates, v_1 recovers the stable semantic direction and suppresses nuisance variation. This can

also be interpreted as a signal-to-noise maximization problem:

$$\mathrm{SNR}(w) = \frac{s^2(w^\top \mu)^2}{w^\top \Sigma_\varepsilon w}, \quad \mathrm{SNR}(v_1) \geq \mathrm{SNR}(w),\ \forall w. \tag{11}$$

Thus, projecting onto v_1 selects the direction that maximizes semantic SNR and minimizes perturbation energy, yielding a provably robust adjustment vector in latent space. Let F denote the local mapping from hidden states to attention outputs, assumed L-Lipschitz. The deviation between using the full perturbation Δh and its rank-1 PCA projection $P_1 \Delta h$ satisfies:

$$\mathbb{E}\|F(h+\beta\Delta h) - F(h+\beta P_1 \Delta h)\| \leq L\beta\sqrt{L_1 \sum_{i=1}^{d} \lambda_i}. \tag{12}$$

Therefore, because L_1 is small (less than 0.07), the influence of discarded components on downstream generation is provably bounded and negligible, mathematically justifying our choice to strictly use $k = 1$ for PCV construction.

D. PCVs for Guiding New Queries

As introduced in our prior work [20], the use of latent vectors to steer ICL can be interpreted through the lens of attention modulation. In a typical multi-head self-attention layer, the attention output for a token x_{query} in the query input X_{query} can be approximated as a weighted combination of its own contextual representation and the latent signal derived from the demonstration input X_{demo}:

$$\mathrm{Att}(x_{\text{query}}W_q, XW_k, XW_v) \approx \alpha h(X_{\text{query}}) + (1-\alpha)h(X_{\text{demo}}), \tag{13}$$

where $\mathrm{Att}(\cdot)$ denotes the attention function, W_q, W_k, W_v are the projection matrices, α is a scalar that represents the sum of normalized attention weights between demonstrations and query examples, and $h(\cdot)$ denotes latent representations. In our method, we replace $h(X_{\text{demo}})$ with our computed PCV to enable compact and training-free contextual modulation:

$$\mathrm{Att}(x_{\text{query}}W_q, XW_k, XW_v) \approx \alpha h(X_{\text{query}}) + (1-\alpha)h_{\text{PCV}}. \tag{14}$$

To apply the PCV during inference, we perform a forward pass on the query and inject the PCV into the hidden states at all token positions across all transformer layers. For each layer $l = 1, 2, \ldots, L$, we compute:

$$\tilde{h}_l = h_l + \beta h^l_{\text{PCV}}, \tag{15}$$

where h_l is the original hidden state, h^l_{PCV} is the l-th layer component of the PCV, and β is a scaling factor that controls the influence of the PCV. We set $\beta = 5$ in this paper. This additive adjustment encourages the model to shift its latent representation in the direction of h^l_{PCV}, which can capture a more stable and perturbation-resilient semantic signal.

To preserve activation scale and avoid distribution shift, we normalize the modified hidden state to match the original ℓ_2 norm:

$$\tilde{h}_l = \tilde{h}_l \cdot \frac{\|h_l\|_2}{\|\tilde{h}_l\|_2}. \tag{16}$$

This normalization ensures that latent injection maintains compatibility with the model's internal representations, allowing stable and semantically aligned report generation under zero-shot or perturbed settings.

IV. Results and Discussion

In this section, we first describe the implementation details. We then report the performance under fully supervised settings on the MIMIC-CXR, CheXpert Plus, and PEIR Gross datasets. Next, we evaluate the zero-shot generation performance under cross-disease and cross-center scenarios, highlighting the adaptability of PCVs without retraining. We further analyze the longitudinal report generation task, which examines temporal reasoning and consistency across follow-up studies. An extensive ablation study is subsequently conducted to examine the contribution of each component and the sensitivity of PCV hyperparameters. Finally, we provide a discussion on the generalizability and scalability of the proposed approach, outlining its potential impact and limitations in broader clinical applications.

A. Implementation Details

In this section, we introduce the experimental datasets, evaluation metrics and training setup.

Datasets: Four publicly available datasets are used to evaluate the generalization capabilities of the proposed PCVs. *IU-Xray* [35] is a widely used benchmark in radiology report generation, consisting of 3,955 reports paired with 7,470 chest X-ray images. Each report may include one or two associated views (frontal and/or lateral). We follow the standard split of 2,069/296/590 for training, validation, and testing, respectively.

MIMIC-CXR [17] is comprised of 368,960 images and 222,758 corresponding reports. Each image–report pair is annotated with 14 disease labels using the CheXpert labeling tool [57]. Following the preprocessing procedure in R2Gen [40], we exclude incomplete studies and obtain a final split of 270,790 training, 2,130 validation, and 3,858 test samples.

Longitudinal-MIMIC [58] is a curated subset of MIMIC-CXR designed for temporal reasoning tasks. It includes 94,169 samples from 26,625 patients with at least two visits. Each sample contains a pair of chest X-rays and associated reports (one from the current visit and one from the previous visit) enabling evaluation of longitudinal report generation. The dataset is split into 92,374 samples for training (26,156 patients), 737 for validation (203 patients), and 2,058 for testing (266 patients).

CheXpert Plus [11] is a large-scale, structured dataset featuring 223,462 chest X-ray and report pairs from 64,725 patients across 187,711 studies. Notably, we pre-trained our R2VLM only on the IU-Xray and MIMIC-CXR datasets. For evaluation on MIMIC-Longitudinal and CheXpert Plus, we adopt a training-free setting to assess the generalization ability of our model. Specifically, for CheXpert Plus, we randomly sample 500 studies from the full dataset to serve as our evaluation subset.

PEIR Gross [36] contains 7,443 images extracted from the Gross collections of 21 PEIR pathology sub-categories. To enable systematic evaluation, we randomly split the dataset into 80/10/10 (train/eval/test) sets. This setup allows us to assess both

the cross-domain generalization and captioning consistency of our proposed PCV framework under a distinct visual modality.

Metrics: We adopt the NLG and clinical efficacy metrics to automatically evaluate the quality of generated medical reports. *NLG Metrics* are used to evaluate the fluency and content similarity between generated reports and ground-truth references. We adopt four widely used metrics: BLEU-N [59], CIDEr [60], ROUGE-L [61], and METEOR [62]. BLEU-N computes the n-gram overlap between generated and reference texts and was originally designed for machine translation. ROUGE-L focuses on the longest common subsequence, capturing sentence-level fluency and structure. METEOR aligns words based on both exact and semantic matches, rewarding recall more heavily and supporting synonym matching. CIDEr, tailored for image captioning, employs TF-IDF weighting to prioritize clinically meaningful terms and downweight common phrases, making it particularly suited to the medical domain.

Clinical Efficacy Metrics evaluate the medical accuracy of generated reports. Specifically, we use the CheXpert labeler [57] to extract 14 diagnostic labels from the generated reports and compare them to ground-truth labels using standard classification metrics: F1 score, Precision, and Recall. Since the IU-Xray dataset lacks CheXpert annotations, these metrics are computed only on datasets that provide ground-truth disease labels.

Training Details: To train our R2VLM backbone, we employ a modular architecture consisting of a frozen MedCLIP-ViT as the visual encoder, a Q-Former as the visual prompt generator, and a frozen LLaMA-2 (7B) as the language decoder. The Q-Former contains 32 learnable queries with an embedding size of 768, followed by a linear projection layer that maps the output to 4096 dimensions to match the input size of LLaMA-2. The total number of trainable parameters is approximately 192 M. We adopt mixed-precision training and pre-train R2VLM for 5 epochs on IU-Xray and 15 epochs on MIMIC-CXR. The model is optimized using Adam with a learning rate of 1×10^{-4} and a weight decay of 0.02. Training a batch of size 8 takes 0.18 seconds on average. All experiments are conducted on 4 NVIDIA RTX A5000 GPUs (24 GB) using Python 3.11.

B. Performances on Fully Supervised Report Generation

Settings: To further assess the utility of our proposed PCVs mechanism, we evaluate its effect in a fully supervised setting. In this setting, we compute the PCVs using 56 randomly selected demonstrations from the training set, where each image is augmented with $K = 10$ perturbed versions via random masking to capture robust latent directions. We report the performances (with respect to the previously defined metrics) of our R2VLM and several state-of-the-art MRG methods for comparison, including R2Gen [40], R2GenCMN [63], PPKED [39], DCL [42], MET [64], R2GPT [10], PromptMRG [13], BiomedGPT [65], [66] and CXP [11].

Supervised Report Generation Performances: As summarized in Tables I, II, and III, the R2VLM model serves as a competitive fully supervised baseline across all three datasets, achieving performance on par with or superior to prior MRG methods. Introducing PCVs leads to consistent and statistically significant improvements over the base models on nearly all metrics. For example, on IU-Xray (Table II), R2VLM+PCVs improves CIDEr from 0.427 to 0.449 ($p < 0.01$), and we observe similar gains across BLEU-4, ROUGE, METEOR and clinical accuracy scores on MIMIC-CXR (Table I), as well as on the PEIR Gross pathology dataset (Table III). These results demonstrate that PCVs provide a non-trivial benefit over MCVs, enhancing both text quality and clinical correctness, while generalizing across model architectures and different imaging modalities. The consistent improvements across three datasets and multiple architectures underscore the plug-and-play nature of PCVs and their ability to stabilize latent representations in a way that generalizes both chest X-rays and pathology images.

Qualitative Analysis: To further evaluate the effectiveness of PCVs, we conduct an analysis on the CXP. As shown in

TABLE I
COMPARISON OF FULLY SUPERVISED MRG PERFORMANCE ON THE MIMIC DATASET. THE BEST AND SECOND PERFORMANCES FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED. STATISTICAL SIGNIFICANCE BETWEEN BASELINE AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN GREEN AND RED INDICATE p-VALUES < 0.05 AND < 0.01, RESPECTIVELY.

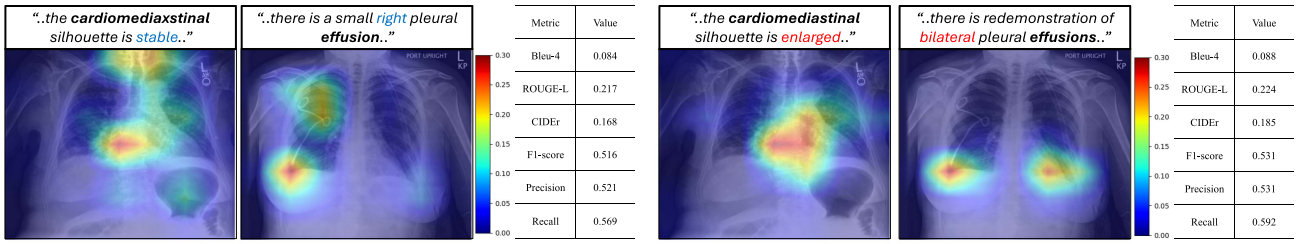
Method	BLEU-4	ROUGE	METEOR	CIDEr	Precision	Recall	F1-score
R2Gen	0.103	0.277	0.142	-	0.333	0.273	0.276
DCL	0.109	0.284	0.150	**0.281**	0.471	0.352	0.373
PromptMRG	0.112	0.268	0.157	-	**0.501**	**0.509**	**0.476**
R2GPT-D	**0.134**	0.297	0.160	0.269	0.392	0.387	0.389
R2VLM	0.128	0.289	0.161	0.265	0.412	0.373	0.395
+ PCVs	0.130	0.291	**0.164**	0.273	0.427	0.385	0.406
CXP	0.108	0.314	0.159	0.246	0.434	0.461	0.437
+ PCVs	0.113	**0.317**	0.162	0.261	0.440	0.475	0.442
BiomedGPT	0.122	0.287	0.159	0.234	0.427	0.413	0.417
+ PCVs	0.124	0.287	0.161	0.240	0.435	0.420	0.428

TABLE II
COMPARISON OF FULLY SUPERVISED MRG PERFORMANCE ACROSS NLG METRICS ON IU-XRAY. THE BEST AND SECOND PERFORMANCES FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED. STATISTICAL SIGNIFICANCE BETWEEN BASELINE AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN GREEN AND RED INDICATE p-VALUES < 0.05 AND < 0.01, RESPECTIVELY.

Method	BLEU-4	ROUGE	METEOR	CIDEr
R2Gen	0.165	0.371	0.187	-
PPKED	0.168	0.376	0.190	0.351
DCL	0.163	**0.383**	0.193	**0.586**
MET	0.172	0.380	0.192	0.435
R2GPT-S	0.156	0.370	0.202	0.405
R2GPT-D	**0.173**	0.377	0.211	0.438
R2VLM	0.168	0.381	0.209	0.427
+PCVs	0.171	0.381	**0.213**	0.449
BiomedGPT-B	0.162	0.285	0.129	0.401
+PCVs	0.166	0.289	0.137	0.437

TABLE III
PERFORMANCES ON THE PEIR GROSS PATHOLOGY DATASET. STATISTICAL SIGNIFICANCE BETWEEN "+MCVS" AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN GREEN AND RED INDICATE p-VALUES < 0.05 AND < 0.01, RESPECTIVELY.

Method	Bleu-4	ROUGE-L	METEOR	CIDEr
R2VLM	0.119	27.9	14.9	32.9
+ MCVs	0.123	28.7	15.1	35.2
+ PCVs	0.125	29.5	15.6	38.0
BiomedGPT-B	0.137	36.0	15.4	122.7
+ MCVs	0.139	36.2	15.7	124.8
+ PCVs	0.140	36.5	16.0	127.5



(a) Generated reports by CXP. (b) Generated reports by CXP + PCVs.

Fig. 3. Illustrations of generated reports by CXP without and with PCVs on the CheXpert Plus dataset. Words in ***bold*** indicate the text tokens explicitly visualized in the cross-attention heatmaps. *Red* words denote positive findings, while those in *blue* represent negative findings. The right-hand tables summarize automatic evaluation metrics.

Fig. 3, we visualize two representative cases comparing the original CXP with its PCV-enhanced counterpart. From the visualizations, we observe that incorporating PCVs enables the model to attend to more clinically relevant image regions in a stable and accurate manner. For instance, in the generation of the term *cardiomediastinal*, the baseline model misplaces its focus, resulting in a generic or incorrect description. In contrast, with PCVs guidance, the model's attention shifts toward the correct anatomical region on the right side of the image, corresponding to the actual location of the cardiomediastinal silhouette, and correctly generates the phrase *enlarged*, aligning with the underlying pathology. In addition to qualitative improvements, the two tables in Fig. 3 demonstrate that PCVs also improve automatic evaluation metrics across both NLG and CE. These gains underscore the ability of PCVs to mitigate cross-modal instability and reinforce semantically grounded generation in VLMs, ultimately enabling more accurate and clinically meaningful medical report generation.

C. Performances on Zero-Shot Report Generation

Settings: To evaluate the zero-shot capabilities of PCVs, we design two clinically inspired scenarios: *cross-center* and *cross-disease* evaluations, using the IU-Xray and MIMIC-CXR datasets.

In the *cross-center* setting, we mimic the deployment of R2VLMs in smaller medical institutions where access to large-scale paired image–text data is limited. Given that reporting styles and terminology vary across institutions, this setting assesses the robustness of PCVs when transferring across different clinical domains. Specifically, we pre-train the R2VLM on IU-Xray and evaluate on MIMIC-CXR, and vice versa, using PCVs to guide zero-shot generation. The *cross-disease* setting reflects scenarios where models must handle rare or emerging diseases absent from the training data. Conducted solely on MIMIC-CXR, this evaluation follows a leave-one-disease-out protocol: in each round, one disease is held out during training and used exclusively for testing. This setup assesses the model's ability to generalize to unseen pathologies without additional disease-specific supervision. All training hyperparameters remain consistent with those used in previous supervised settings.

Cross-center Performance: Tables IV and V present the results of cross-center zero-shot report generation on the IU-Xray and MIMIC-CXR datasets. These experiments simulate real-world deployment scenarios, where a model trained in one clinical center is applied to another without fine-tuning. We report both NLG metrics and CE metrics to assess generalization ability.

Across all baseline models, incorporating MCVs consistently improves performance in both directions. For instance, in the IU-Xray $\rightarrow$ MIMIC-CXR setting, R2GPT-D improves its CIDEr score from 0.072 to 0.134 with MCVs. Similar gains are observed in BLEU-4 (from 0.071 to 0.101), ROUGE (0.218 to 0.251), and METEOR (0.113 to 0.122). In the reverse direction, MIMIC-CXR $\rightarrow$ IU-Xray, R2GPT-D+MCVs further improves BLEU-4 from 0.083 to 0.112 and METEOR from 0.163 to 0.177. Our own Med-LMM (R2VLM) shows the strongest gains, with MCVGen achieving a CIDEr of 0.142 and METEOR of 0.128, substantially outperforming baseline results. Upon further integrating PCVs, we observe additional performance gains across nearly all settings. This is particularly evident in CIDEr, where PCVs help R2VLM improve from 0.142 to 0.157in the IU $\rightarrow$ MIMIC direction, and from 0.280 to 0.288 in the reverse. The improvements stem from the fact that PCVs guide the model's latent states toward more stable and semantically consistent directions, effectively reducing the influence of perturbation-sensitive visual features. These improvements are statistically significant in most cases, especially for BLEU-4, CIDEr, and F1-score metrics. Clinical efficacy metrics in Table V reinforce this trend. For example, R2VLM+PCVs achieves an F1-score of 0.258, surpassing both the base model (0.208) and the MCV-only variant (0.237). Similar improvements are observed with R2GPT-D, where the F1-score increases from 0.197 to 0.263 after applying PCVs. Moreover, statistical significance further confirms that PCVs consistently outperform MCVs with $p < 0.01$, demonstrating stronger robustness and cross-domain generalization. We additionally conducted a blinded human evaluation with three board-certified radiologists assessing 100 randomly selected cases. On average, the raters preferred the PCV-generated reports in 52.3% of the cases (Table VII), confirming their higher clinical faithfulness and coherence. To quantify inter-rater reliability, we computed the Fleiss' kappa statistic, which yielded a score of 0.73, indicating a "Substantial

TABLE IV
COMPARISON OF CROSS-CENTER ZERO-SHOT LEARNING PERFORMANCE ACROSS NLG METRICS ON THE IU-XRAY AND MIMIC DATASET. THE DATASETS ON THE LEFT AND RIGHT OF -> INDICATE THE PRE-TRAINING AND TESTING SETS, RESPECTIVELY. STATISTICAL SIGNIFICANCE BETWEEN "+MCVS" AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN GREEN AND RED INDICATE p-VALUES < 0.05 AND < 0.01, RESPECTIVELY.

Method	IU-Xray->MIMIC-CXR				Method	MIMIC-CXR->IU-Xray			
	BLEU-4	ROUGE	METEOR	CIDEr		BLEU-4	ROUGE	METEOR	CIDEr
R2Gen	0.067	0.201	0.108	0.072	R2Gen	0.059	0.253	0.131	-
R2GenCMN	0.061	0.213	0.105	0.069	DCL	0.074	0.267	0.152	-
DCL	0.064	0.208	0.105	0.063	PromptMRG	0.098	0.281	0.160	-
R2GPT-S	0.064	0.211	0.108	0.066	R2GPT-S	0.051	0.229	0.114	0.196
+ MCVs	0.088	0.247	0.118	0.083	+ MCVs	0.083	0.291	0.149	0.207
+ PCVs	0.091	0.247	0.121	0.105	+ PCVs	0.089	0.304	0.152	0.215
R2GPT-D	0.071	0.218	0.113	0.072	R2GPT-D	0.083	0.275	0.163	0.231
+ MCVs	0.101	0.251	0.122	0.134	+ MCVs	0.112	0.308	0.177	0.236
+ PCVs	0.105	0.253	0.124	0.153	+ PCVs	0.116	0.308	0.184	0.249
R2VLM	0.068	0.215	0.115	0.069	R2VLM	0.087	0.279	0.137	0.252
+ MCVs	0.106	0.256	0.128	0.142	+ MCVs	0.125	0.313	0.185	0.280
+ PCVs	0.108	0.259	0.132	0.157	+ PCVs	0.127	0.315	0.190	0.288

TABLE V
COMPARISON OF CROSS-CENTER ZERO-SHOT LEARNING PERFORMANCE ACROSS CE METRICS IN THE IU-XRAY TO MIMIC-CXR ADAPTATION. STATISTICAL SIGNIFICANCE BETWEEN "+MCVS" AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN RED INDICATE p-VALUES < 0.01.

Method	Precision	Recall	F1-score
R2Gen	0.139	0.134	0.135
R2GenCMN	0.145	0.132	0.134
DCL	0.156	0.143	0.146
R2GPT-S	0.128	0.116	0.119
+ MCVs	0.181	0.169	0.173
+ PCVs	0.194	0.201	0.193
R2GPT-D	0.208	0.193	0.197
+ MCVs	0.242	0.226	0.234
+ PCVs	0.261	0.245	0.263
R2VLM	0.211	0.206	0.208
+ MCVs	0.249	0.235	0.237
+ PCVs	0.263	0.241	0.258

TABLE VI
COMPARISON OF CROSS-DISEASE ZERO-SHOT LEARNING PERFORMANCE ON THE MIMIC DATASET. STATISTICAL SIGNIFICANCE BETWEEN "+MCVS" AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN GREEN AND RED INDICATE p-VALUES < 0.05 AND < 0.01, RESPECTIVELY.

Method	BLEU-4	METEOR	ROUGE	CIDEr	Precision	Recall	F1-score
R2GPT-S	0.098	0.253	0.122	0.093	0.173	0.161	0.163
+ MCVs	0.105	0.267	0.132	0.121	0.192	0.179	0.180
+ PCVs	0.110	0.271	0.134	0.135	0.198	0.188	0.195
R2GPT-D	0.117	0.277	0.136	0.144	0.253	0.238	0.244
+ MCVs	0.124	0.285	0.145	0.187	0.289	0.267	0.275
+ PCVs	0.126	0.291	0.145	0.191	0.293	0.277	0.293
R2VLM	0.113	0.269	0.141	0.140	0.261	0.232	0.246
+ MCVs	0.129	0.291	0.148	0.203	0.294	0.277	0.281
+ PCVs	0.130	0.295	0.152	0.206	0.308	0.278	0.305

Agreement" among the clinicians. However, qualitative analysis of edge cases reveals a fundamental subjectivity in clinical evaluation: while radiologists generally favor PCV-enhanced reports for their superior diagnostic sensitivity, individuals weigh the severity of false-negative omissions against the risks of minor false-positive hallucinations and syntactic flaws differently. This underscores that although PCVs significantly improve factual grounding, perfectly balancing detailed clinical sensitivity with flawless language generation remains a complex challenge, reinforcing the necessity of human-in-the-loop deployment. More details are presented in the Appendix, available online.

TABLE VII
RADIOLOGIST EVALUATION OF GENERATED REPORTS ON 100 MIMIC-CXR TEST CASES. EACH ENTRY INDICATES THE PERCENTAGE OF CASES FOR WHICH REPORTS FROM EACH MODEL WERE PREFERRED AS THE MOST CLINICALLY ACCURATE.

Rater	R2VLM	+MCVs	+PCVs
R1 (2 yrs exp.)	21%	28%	**51%**
R2 (3 yrs exp.)	25%	24%	**51%**
R3 (Professor, 15+ yrs)	18%	27%	**55%**

In summary, these results demonstrate that MCVGEN establishes state-of-the-art zero-shot MRG performance across both datasets. The added benefit from PCVs highlights their role in stabilizing multimodal alignment, suppressing perturbation-induced semantic drift, and enhancing the model's ability to generalize across institutions and writing styles.

Cross-disease Performance: In our cross-disease evaluation, we meticulously maintained a balanced distribution by utilizing 14 demonstrations with equal representation across diseases. This strategy ensures that our model captures multi-modal contextual information from categories it has not previously learned. By employing this approach, we aimed to enhance the model's generalizability and robustness across various unencountered medical conditions, thereby demonstrating its efficacy in handling diverse diagnostic scenarios to describe novel diseases. Table VI reports the results of the cross-disease zero-shot evaluation on the MIMIC-CXR dataset. Across all models, integrating MCVs yields consistent gains. For example, R2GPT-S improves in BLEU-4 from 0.098 to 0.105 and in CIDEr from 0.093 to 0.121. R2GPT-D shows stronger improvements, with BLEU-4 rising from 0.117 to 0.124 and CIDEr from 0.144 to 0.187. Our proposed Med-LMM (R2VLM) benefits the most: incorporating MCVs increases BLEU-4 from 0.113 to 0.129 and CIDEr from 0.140 to 0.203.

More importantly, we observe significant improvements in CE metrics accroding to the significance statistic. For instance, R2VLM+MCVs boosts F1-score from 0.246 to 0.281, with consistent gains in both precision and recall. R2GPT-D+MCVs achieves a precision of 0.289 and F1-score of 0.275. These improvements highlight the ability of MCV to improve clinical

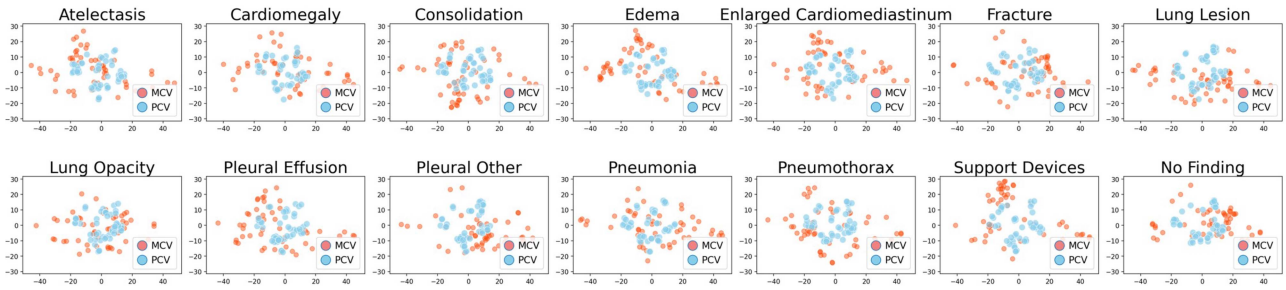


Fig. 4. t-SNE visualization of MCV and PCV latent representations. Each subfigure contains samples labeled with the corresponding chest X-ray disease category (covering all 14 labels).

factual correctness in unseen disease conditions, critical for real-world applicability. When further incorporating PCVs, we observe additional performance gains across nearly all models. For R2VLM, the F1-score reaches 0.305, while CIDEr climbs to 0.206. These enhancements demonstrate that PCVs help align the model's latent representations along stable and robust semantic directions, mitigating the influence of noisy or biased visual cues. This leads to more clinically reliable report generation even in insufficient data or rare disease scenarios, enabling medical VLMs to generalize to new diseases with minimal supervision.

We further visualize and compare the latent distributions of MCVs and PCVs across 14 chest X-ray disease labels in Fig. 4. While the MCV embeddings form broad and loosely mixed clusters, PCV embeddings exhibit more compact and well-separated structures aligned with semantic disease categories. This suggests that projecting perturbation-induced latent shifts onto the principal semantic direction reduces dispersion in the latent space and improves class-level organization. Notably, diseases with visually similar manifestations (e.g., atelectasis, effusion, consolidation) show reduced overlap under PCV embeddings, indicating improved intra-class consistency and inter-class separability. These qualitative trends complement the quantitative robustness gains reported in the main experiments.

D. Performances on Longitudinal Report Generation

Settings: Longitudinal Report Generation (LRG) refers to the task of generating follow-up radiology reports that not only describe the current image but also contextualize it in light of prior studies. This task is clinically important, as it mirrors real-world radiological workflows where disease progression, treatment response, and anatomical changes are evaluated over time. We conducted experiments on the Longitudinal-MIMIC dataset [58]. We trained two representative models, R2GPT-D and our proposed R2VLM, using paired current and previous CXR-report demonstrations. During inference, we applied PCVs computed from training demonstrations to guide generation in a training-free manner. We also compare our models with four STOA LRG-specific approaches, Prefilling [58], HC-LLM [22], MLRG [67] and HERGen [21], respectively.

Qualitative and Quantitative Analysis on the Longitudinal-MIMIC: Table VIII summarizes the performance of various fully supervised approaches on the Longitudinal-MIMIC dataset.

TABLE VIII
COMPARISON OF FULLY SUPERVISED LRG APPROACHES ON THE LONGITUDINAL-MIMIC DATASET. STATISTICAL SIGNIFICANCE BETWEEN BASELINE AND "+PCVS" USING PAIRED t-TESTS ARE REPORTED, WHERE NUMBERS IN GREEN AND RED INDICATE p-VALUES < 0.05 AND < 0.01, RESPECTIVELY.

Method	BLEU-4	METEOR	ROUGE	Precision	Recall	F1-score
Prefilling	0.099	0.137	0.271	0.538	0.434	0.480
HC-LLM	0128	0.160	0.287	0.417	0.357	0.357
MLRG	0.158	0.176	0.320	0.549	0.468	0.505
HERGen	0.122	0.256	0.285	-	-	-
R2GPT-D	0.115	0.228	0.274	0.402	0.355	0.350
+ PCVs	0.119	0.243	0.287	0.406	0.365	0.371
R2VLM	0.113	0.216	0.266	0.385	0.361	0.366
+ PCVs	0.118	0.235	0.280	0.397	0.368	0.372

Among existing LRG-specific baselines, **MLRG** and **Prefilling** achieve strong results in both NLG and CE metrics, with MLRG reaching the highest F1-score (0.505) and ROUGE (0.320), while Prefilling demonstrates robust clinical grounding with a precision of 0.538. Our proposed R2VLM+PCVs achieves competitive results without task-specific architectural design or retraining. Specifically, R2VLM+PCVs outperforms its base model across all metrics and attains an F1-score of 0.372 and NLG-aligned improvements (e.g., METEOR from 0.216 to 0.235), rivaling the performance of models tailored for longitudinal reasoning. Notably, both R2GPT-D and R2VLM benefit from the inclusion of PCVs, which consistently improve report fluency, clinical accuracy, and temporal coherence. This demonstrates that PCVs effectively inject stable, semantically meaningful signals into the model's latent space, mitigating the impact of visual variation across time and supporting more faithful disease progression tracking. In summary, although our framework is not explicitly designed for longitudinal tasks, the results show that PCVs enable general-purpose VLMs to perform competitively with dedicated LRG models, highlighting their adaptability and robustness in complex, temporally structured clinical settings.

Fig. 5 presents the generated reports before and after incorporating PCVs. Notably, the baseline model fails to correctly estimate the position of the endotracheal tube and misses important clinical findings such as heart enlargement. In contrast, after applying PCVs, the model produces a more precise description of the tube tip location "approximately 3.5 cm above the carina" and correctly identifies the cardiomediastinal silhouette

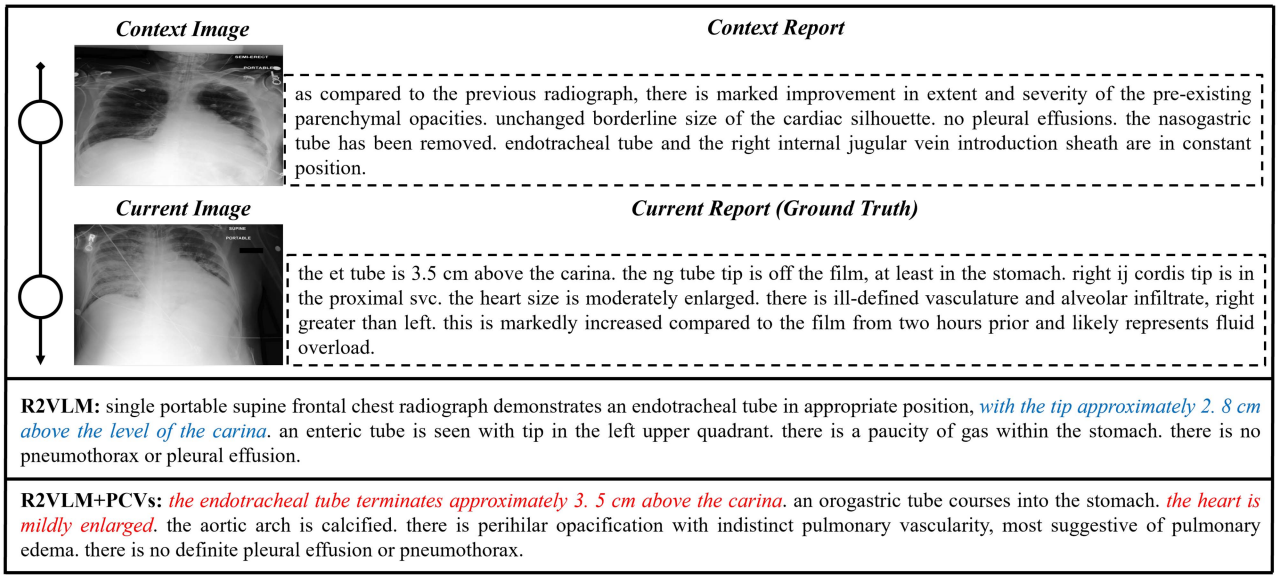


Fig. 5. Illustrations of generated longitudinal report by R2VLM and our R2VLM+PCVs. *Red* words denote positive findings, while those in *blue* represent negative findings.

TABLE IX
COMPARISON OF DIFFERENT SETTINGS TO GENERATE AND APPLY MCV IN THE IU-XRAY TO MIMIC ADAPTATION. THE SETTING "A+C" REPRESENTS THE FINAL PCV.

Setting	BLEU-4	CIDEr	Precision	Recall	F1-Score
a+c	0.108	0.157	0.263	0.241	0.258
b+c	0.071	0.102	0.221	0.198	0.203
a+d	0.102	0.137	0.253	0.236	0.238
a+e	0.096	0.124	0.218	0.193	0.195

as "mildly enlarged." This case demonstrates the effectiveness of PCVs in stabilizing the model's visual representation across temporally spaced studies, enabling more accurate and clinically grounded longitudinal report generation.

E. Ablation Study for PCV

Different settings for generating and applying PCVs: To evaluate the design choices in PCVs, we conduct a series of ablation studies exploring alternative strategies for generating and applying PCVs and present results in Table IX. For PCVs generation, we compare our averaging-based approach (setting a) with a recurrent formulation (setting b), where the MCV from each demonstration is used to shift the latent state of the subsequent demonstration. This recurrent MCV (b+c) results in significantly worse performance, achieving only 0.071 BLEU-4 and 0.203 F1-score. Regarding PCVs application, our default setting (a+c) injects the averaged PCV into all token positions of the query across all layers. We compare this to (a+d), where PCVs are added to the first token position only. While MCVs are most effective when injected only at the first token position, as they primarily serve as task-level guidance. Our results suggest that PCVs require broader integration. Since PCVs encode stable features that mitigate representational drift in multimodal embeddings, applying them across all token positions proves necessary to fully realize their effect. This contrast highlights the distinct roles of MCVs and PCVs in guiding generation: one for semantic alignment, the other for representation stabilization. Finally, we assess a replacement strategy (a+e), where the first token's hidden state is entirely replaced by the PCV, inspired by Task Vectors [50]. This approach underperforms across all metrics (e.g., CIDEr 0.124 vs. 0.157, F1-score 0.195 vs. 0.258), indicating that preserving the original query context while gently shifting it in the direction of the PCV is more effective than hard replacement.

Demonstration Selection: We conducted an ablation study to investigate how the selection and number of demonstrations influence zero-shot performance in the IU-Xray → MIMIC adaptation. As shown in Table XI, we compare randomly selected demonstrations (14-rd) with equally distributed demonstrations (14-eq), where each of the 14 disease categories in MIMIC is represented once. The results show that balanced coverage of disease categories significantly improves performance: 14-eq outperforms 14-rd across all metrics, boosting BLEU-4 from 0.071 to 0.094 and F1-score from 0.201 to 0.237. This suggests that semantic diversity among demonstrations plays a more important role than sample quantity alone. We further tested increasing the number of demonstrations (28, 56, and 112), all equally distributed across categories. While minor improvements are observed (e.g., BLEU-4 rising to 0.108 and F1-score to 0.258 with 56 samples), the performance saturates quickly. Notably, the difference between 56 and 112 demonstrations is negligible, indicating diminishing returns. Based on these findings, we adopt 56 equally distributed demonstrations in all subsequent evaluations to balance performance and efficiency.

Multi-modal Intervention: Table X presents an ablation study comparing visual (**v**), textual (**t**), and joint (**v+t**) interventions

TABLE X
COMPARISON OF FULLY SUPERVISED MRG PERFORMANCE ON THE MIMIC DATASETS. THE HIGHEST PERFORMANCES FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD**. "V" AND "T" REFER TO VISUAL AND TEXTUAL INTERVENTION, RESPECTIVELY.

Method	BLEU-4	ROUGE	METEOR	CIDEr	Precision	Recall	F1-score
R2Gen	0.103	0.277	0.142	-	0.333	0.273	0.276
DCL	0.109	0.284	0.150	**0.281**	0.471	0.352	0.373
PromptMRG	0.112	0.268	0.157	-	**0.501**	**0.509**	**0.476**
R2GPT-D	**0.134**	0.297	0.160	0.269	0.392	0.387	0.389
CXP	0.108	0.314	0.159	0.246	0.434	0.461	0.437
+PCVs (v)	0.113	**0.317**	0.162	0.261	0.440	0.475	0.442
+PCVs (t)	0.111	0.314	0.158	0.252	0.438	0.466	0.438
+PCVs (v+t)	0.113	0.316	0.160	0.261	0.440	0.472	0.441
R2VLM	0.128	0.289	0.161	0.265	0.412	0.373	0.395
+PCVs (v)	0.130	0.291	**0.164**	0.273	0.427	0.385	0.406
+PCVs (t)	0.130	0.289	0.163	0.268	0.422	0.380	0.403
+PCVs (v+t)	0.130	0.291	**0.164**	0.271	0.425	0.383	0.406

TABLE XI
ABLATION STUDY OF DEMONSTRATION SELECTION IN THE IU-XRAY TO MIMIC ADAPTATION ACROSS BOTH NLG AND CE METRICS, WHERE *"RD"* AND *"ED"* REFER TO RANDOMLY DISTRIBUTED AND EQUALLY DISTRIBUTED. THE BEST AND SECOND PERFORMANCES FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED.

Setting	BLEU-4	ROUGE	METOER	CIDEr	F1-score
14 (rd)	0.071	0.219	0.106	0.073	0.201
14 (eq)	0.094	0.243	0.128	0.142	0.237
28 (eq)	0.103	0.257	0.126	0.148	0.242
56 (eq)	**0.108**	0.259	0.132	**0.157**	**0.258**
112 (eq)	**0.108**	**0.261**	**0.134**	0.155	0.253

TABLE XII
ABLATION STUDY OF THE HYPERPARAMETER β CONTROLLING THE QUALITY OF LATENT MODULATION DURING PCV INJECTION. RESULTS ARE REPORTED FOR THE IU-XRAY $\rightarrow$ MIMIC-CXR ADAPTATION USING THE R2VLM BACKBONE.

β	NLG Metrics		Clinical Efficiency (CE) Metrics		
	BLEU-4	CIDEr	Precision	Recall	F1-score
1	0.104	0.140	0.245	0.233	0.235
3	0.106	0.147	0.256	0.238	0.245
5 (default)	**0.108**	**0.157**	**0.263**	**0.241**	**0.258**
10	0.106	0.149	0.261	0.238	0.252

under fully supervised settings. Consistent with our earlier analysis, visual perturbation–based PCVs provide the most significant improvements across both NLG and CE metrics, confirming that report generation quality is primarily constrained by visual stability. Specifically, on MIMIC-CXR, R2VLM+PCVs (v) improves BLEU-4 from 0.128 to 0.130 and METEOR from 0.161 to 0.164, while textual intervention alone yields marginal gains. Combining both modalities (v+t) achieves performance comparable to visual-only guidance, suggesting that excessive textual perturbation may not further enhance robustness. These results validate our design choice to focus on perturbation-aware visual stabilization as the dominant factor influencing multimodal report generation accuracy and clinical fidelity.

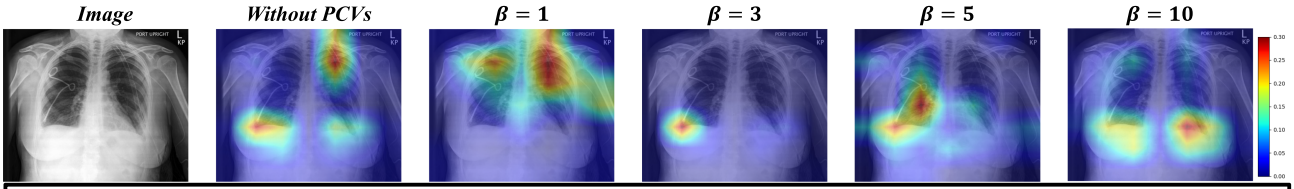
Effect of varying β: To evaluate its sensitivity, we conducted both quantitative and qualitative analyses by varying $\beta \in \{1, 3, 5, 10\}$ on the cross-center IU-Xray $\rightarrow$ MIMIC-CXR task using the R2VLM backbone. As summarized in Table XII, performance steadily improves as β increases from 1 to 5, but slightly decreases at $\beta = 10$, indicating that overly strong modulation can distort latent activations and reduce fluency. We therefore fix $\beta = 5$ for all experiments, which achieves the best trade-off between robustness and stability. Furthermore, the visualization in the manuscript provides qualitative evidence consistent with this trend. When β is too small (e.g., $\beta = 1$), the influence of PCVs is marginal, leading to attention patterns similar to the base model and weak alignment with pathological regions. As β increases, the model exhibits stronger and more localized attention on clinically relevant structures. At $\beta = 5$, the model achieves a balanced focus—accurately attending to regions indicative of pneumothorax and generating semantically consistent diagnostic phrases that match CheXpert labels. Furthermore, as shown in Fig. 6, when $\beta = 10$ attention becomes scattered and textual coherence declines, occasionally producing hallucinated terms. Importantly, since all tasks in this study share the same imaging modality and latent activation scale, β generalizes well across datasets and model variants without the need for re-tuning.

Noise-based Visual Perturbation: To investigate the effect of different perturbation strategies on the construction of PCVs, we conducted a comparative study evaluating both masking-based and noise-based perturbations. Masking has been our primary choice, as it directly simulates localized information loss—one of the most common causes of representational drift in vision-language models. However, to further assess robustness, we additionally experimented with Gaussian noise injection. Specifically, for each image I, we generated a noisy counterpart as:

$$I_{\text{noisy}} = \lambda I + (1 - \lambda)\mathcal{N}(0, 1),$$

where $\lambda \in [0, 1]$ controls the perturbation strength. We evaluated multiple λ values to analyze the sensitivity of PCV stability to noise levels. As shown in Fig. 7, performance progressively degrades as λ decreases, indicating that strong global noise disrupts semantic alignment and latent feature consistency. When $\lambda \geq 0.8$, the results become comparable to those of the masking-based PCV, suggesting that mild perturbations are sufficient to capture stable latent directions, while excessive noise introduces distributional distortions that impair semantic coherence. We attribute this difference to the nature of the perturbations: uniform Gaussian noise alters the overall intensity distribution of the image, affecting all spatial regions equally and thereby disturbing the global visual context. In contrast, masking affects only localized patches, preserving the global anatomical structure and enabling the model to focus on reconstructing semantically meaningful features. Consequently, masking remains the most effective and interpretable strategy for constructing perturbation-aware and semantically consistent PCVs.

Scalability Across Model Sizes: To evaluate the scalability and generalizability, we conduct an additional scaling-up study on the CheXpertPlus dataset. We assess the plug-and-play effectiveness of our approach across a wide spectrum of foundation model parameters, specifically utilizing the BioMedGPT [65] family (33 M, 93 M, and 182 M) and the MedGemma [68] family (4B and 27B). As illustrated in Fig. 8, the results reveal two findings. First, scaling up the model capacity inherently improves the baseline clinical factual consistency. However, the marginal benefit of this brute-force scaling is relatively limited given



Ground Truth: Left upper extremity PICC line, left chest tube, and right pigtail pleural drain are again seen. A small right sided pneumothorax has increased in size compared with the prior examination. There is redemonstration of small bilateral pleural effusions. [pneumothorax, pleural effusion, support devices]
R2VLM: there is a left picc line with the tip at the cavoatrial junction. there is a left chest tube. there is a small right apical pneumothorax. there is a small right pleural effusion. there is a small left pleural effusion. there is no focal consolidation. [consolidation, pneumothorax, pleural effusion, support devices]
R2VLM+PCVs ($\beta = 1$): there is a left chest tube in place. no pneumothorax is identified. there is a small right pleural effusion. there is mild bibasilar atelectasis. there is a left upper extremity picc line with the tip in the superior vena cava. [atelectasis, pneumothorax, pleural effusion, support devices]
R2VLM+PCVs ($\beta = 3$): there is a left chest tube in place with a small left apical pneumothorax. no significant change in the appearance of the left upper extremity picc line. there is persistent right pleural effusion. [pneumothorax, pleural effusion, support devices]
R2VLM+PCVs ($\beta = 5$): there is a left upper extremity picc line with the tip seen at the cavoatrial junction. there is a left chest tube, no pneumothorax. there is a small right apical pneumothorax. small bilateral pleural effusions are present. [pneumothorax, pleural effusion, support devices]
R2VLM+PCVs ($\beta = 10$): there is a small right apical pneumothorax. no pulmonary nodules. the cardiomediastinal silhouette is within normal limits the lungs. there is a right pleural effusion. [enlarged cardiomediastinum, lung lesion, pneumothorax, pleural effusion]

Fig. 6. Effect of different β values on report generation using PCVs. We visualize the cross-attention heatmaps under varying levels of PCV influence. Each row shows the same input image, and each column corresponds to a different β setting. The generated report is displayed below each cross-attention heatmap when generating the word *pneumothorax*, with the CheXpert labeler's output appended in brackets. Red tokens indicate positive findings, while blue tokens indicate negative findings.

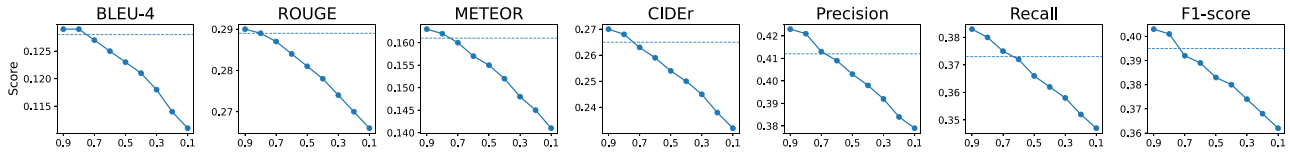


Fig. 7. Ablation study on Gaussian noise perturbation (λ) for constructing PCVs. The blue dashed lines denote the baseline R2VLM performance without PCVs.

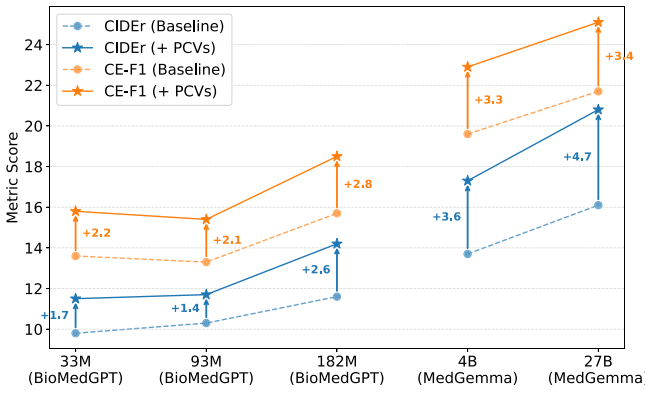


Fig. 8. Performance scaling and PCV improvements across different model sizes on the CheXpert Plus dataset.

the computational cost; for example, scaling the parameters by over 140 times (from BioMedGPT-182 M to MedGemma-27B) yields a baseline CE-F1 improvement from 15.7 to 21.7. Second, and more importantly, integrating our training-free PCVs consistently provides a substantial and orthogonal performance margin over the vanilla baselines across all model scales. The increment value of PCVs remains remarkably stable: it yields a +2.8 CE-F1 improvement on the lightweight 182 M model and an even larger +3.4 improvement on the massive 27B model. These findings empirically demonstrate that while scaling up foundation models is useful, it yields diminishing returns for mitigating hallucinations. In contrast, PCVs effectively bridges this gap and grounds the generation without any parameter updates.

Sensitivity Analysis over the Number of Perturbations: To empirically justify our default choice of $P = 5$ for extracting the PCV, we conduct a comprehensive sensitivity analysis over the number of perturbations. Specifically, we vary $P \in \{1, 2, 3, 5, 10, 30, 50\}$ and evaluate the CIDEr and CE-F1 scores, alongside the corresponding computational overhead (preprocessing time per sample in milliseconds). The results are illustrated in Fig. 9, which plots the trajectory up to $P = 10$;

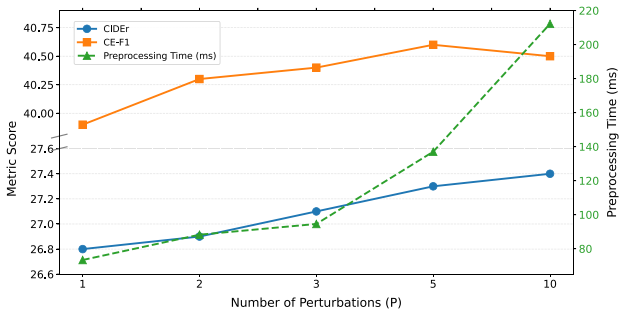


Fig. 9. Sensitivity analysis over the number of perturbations (P) for PCV construction.

we omit plotting higher values because the performance metrics plateau while the processing time explodes. As P increases from 1 to 5, we observe a steady improvement in both CIDEr (from 26.8 to 27.3) and CE-F1 (from 39.9 to 40.6). This confirms our theoretical intuition that aggregating multiple perturbation-induced shifts allows PCA to filter out random high-frequency noise and identify a more robust, invariant semantic direction. While further increasing the number of perturbations to $P = 10$ yields a marginal improvement in CIDEr (from 27.3 to 27.4), it simultaneously slightly reduces the clinical efficacy score (CE-F1 drops from 40.6 to 40.5). More importantly, the preprocessing time scales more than linearly with P; moving from $P = 5$ to $P = 10$ increases the computational overhead by approximately 55% (from 137ms to 212ms). Although not plotted, setting P to extreme values like 30 or 50 further exacerbates this computational bottleneck (requiring 570ms and over 1,040ms per image, respectively) without providing any proportional clinical benefits. Therefore, these results suggest that $P = 5$ offers a practical balance between estimator stability and computational efficiency.

F. Discussion

The proposed PCV framework represents a promising step toward enabling efficient and reliable use of VLMs in MRG. As foundation models continue to grow in scale and generalization capabilities, their application to the medical domain is becoming increasingly attractive. In particular, the emergence of VLMs has opened new opportunities for unifying visual and textual reasoning, offering a natural fit for tasks such as MRG. Looking forward, it is foreseeable that domain-specific VLMs tailored for healthcare will become a critical tool in both research and clinical practice. However, while access to pre-trained VLMs is becoming more feasible for many researchers, the computational cost and data requirements of fine-tuning these models remain prohibitively high, especially in the medical domain, where data is scarce, sensitive, and expensive to annotate. This presents a pressing need for training-free adaptation strategies that allow existing VLMs to be efficiently deployed in new settings. Our work addresses this need by introducing PCVs, which enable latent guidance of VLMs through compact representations derived from a small number of demonstrations, without requiring any model parameter updates.

The proposed method addresses two critical bottlenecks in deploying medical VLMs: adapting to domain shifts with minimal supervision and mitigating persistent generative hallucinations. In real-world clinical environments, institutions frequently face variations in disease distributions, reporting styles, and imaging protocols. While off-the-shelf, general-domain VLMs encode broad knowledge from their pretraining corpora, they typically lack the rigorous clinical grounding required for high-stakes tasks like MRG. Standard ICL can activate latent knowledge but does not endow the model with new, robust understanding, making it insufficient for achieving clinically reliable performance. Our approach overcomes this by utilizing principal component analysis to extract stable, perturbation-invariant semantic directions from visual demonstrations. Rather than retraining the network, PCVs steer the VLM's internal representations to filter out noisy visual signals, effectively guiding the generation trajectory toward clinically accurate and structurally consistent outputs.

Regarding limitations and future directions, it is crucial to explicitly bound the scope of our current empirical claims. First, while our cross-modality experiments on the PEIR Gross dataset successfully demonstrate that PCVs can adapt to domain shifts—specifically the transition from radiological X-rays to macroscopic gross pathology photographs—these results do not evidence generalizability to 3D volumetric modalities such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), nor do they validate efficacy on non-English medical reports. Extending our training-free latent steering to complex volumetric data and multilingual settings remains a vital direction for future work. Second, while our expanded scaling-up experiments demonstrate that PCVs consistently improve zero-shot transfer performance across a wide spectrum of foundation models (up to 27B parameters), we acknowledge the constraints imposed by computational resources. Specifically, we were unable to perform fully supervised fine-tuning on massive foundation models like MedGemma-27B. Consequently, the empirical performance gap between our training-free, PCV-enhanced zero-shot approach and a fully supervised, task-specific fine-tuned massive VLM remains unexplored. Future work will investigate this gap to determine the absolute upper bound of clinical efficacy when combining parameter-efficient fine-tuning with PCV-guided latent steering at the tens-of-billions parameter scale.

V. Conclusion

In this work, we introduce PCVs, a training-free and robust framework for guiding VLMs in MRG. Our approach addresses two key challenges in practical adaption of VLMs to clinical settings: the high computational cost of fine-tuning on domain-specific data, and the instability of visual-language representations that often leads to hallucinations. By extracting MCVs from multimodal demonstrations and applying principal component analysis, PCVs capture stable semantic directions that can steer generation toward clinically accurate outputs. Extensive experiments on IU-Xray, MIMIC-CXR, CheXpert Plus, Peir Gross and Longitudinal-MIMIC datasets demonstrate

that PCVs significantly improve performance in both zero-shot and supervised settings. Notably, our method enables rapid adaptation to rare diseases and clinical domains without retraining, offering a lightweight and scalable solution for medical AI. This is especially valuable in resource-limited environments, where the shortage of radiologists and annotated data often impedes timely diagnosis. By reducing dependence on labeled data and increasing the robustness of VLMs, PCVs provide a practical pathway toward more accessible and trustworthy AI-assisted healthcare.

REFERENCES

[1] Z. Chen, Y. Bie, H. Jin, and H. Chen, "Large language model with region-guided referring and grounding for CT report generation," *IEEE Trans. Med. Imag.*, vol. 44, no. 8, pp. 3139–3150, Aug. 2025.

[2] M. Li, W. Cai, K. Verspoor, S. Pan, X. Liang, and X. Chang, "Cross-modal clinical graph transformer for ophthalmic report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20656–20665.

[3] V. M. Rao et al., "Multimodal generative AI for medical image interpretation," *Nature*, vol. 639, no. 8056, pp. 888–896, 2025.

[4] OpenAI, "GPT-4o technical report," 2024, Accessed: May 2025. [Online]. Available: https://openai.com/index/gpt-4o

[5] H. Liu, C. Li, Y. Wang, Z. Gan, and J. Gao, "Visual instruction tuning," 2023, *arXiv:2304.08485*.

[6] G. Team et al., "Gemini: A family of highly capable multimodal models," 2023, *arXiv:2312.11805*.

[7] J. Pan et al., "MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMS) via reinforcement learning," 2025, *arXiv:2502.19634*.

[8] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 7433–7442, doi: 10.1109/CVPR52729.2023.00718.

[9] Z. Wang, L. Wang, X. Li, and L. Zhou, "Diagnostic captioning by cooperative task interactions and sample-graph consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 6585–6598, Aug. 2025.

[10] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen LLMs," *Meta- Radiol.*, vol. 1, no. 3, 2023, Art. no. 100033.

[11] P. Chambon et al., "CheXpert plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats," 2024, *arXiv:2405.19538*.

[12] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[13] H. Jin, H. Che, Y. Lin, and H. Chen, "PromptMRG: Diagnosis-driven prompts for medical report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 2607–2615.

[14] M. Li et al., "Contrastive learning with counterfactual explanations for radiology report generation," 2024, *arXiv:2407.14474*.

[15] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2577–2586.

[16] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," *World Wide Web*, vol. 26, no. 1, pp. 253–270, 2023.

[17] A. E. W. Johnson et al., "MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.

[18] M. Li et al., "FFA-IR: Towards an explainable and reliable medical report generation benchmark," in *Proc. Adv. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021.

[19] X. Zhang, J. N. Acosta, J. Miller, O. Huang, and P. Rajpurkar, "ReXGradient-160K: A large-scale publicly available dataset of chest radiographs with free-text reports," 2025, *arXiv:2505.00228*.

[20] R. Liu, M. Li, S. Zhao, L. Chen, X. Chang, and L. Yao, "In-context learning for zero-shot medical report generation," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 8721–8730.

[21] F. Wang, S. Du, and L. Yu, "HERGen: Elevating radiology report generation with longitudinal data," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 183–200.

[22] T. Liu et al., "HC-LLM: Historical-constrained large language models for radiology report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 5595–5603.

[23] Q. Dong et al., "A survey on in-context learning," 2022, *arXiv:2301.00234*.

[24] J. Wei et al., "Finetuned language models are zero-shot learners," in *Proc. Int. Conf. Learn. Representations*, 2021.

[25] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.

[26] T. Tu et al., "Towards conversational diagnostic artificial intelligence," *Nature*, vol. 642, pp. 442–450, 2025.

[27] S. Liu, L. Xing, and J. Zou, "In-context vectors: Making in context learning more effective and controllable through latent space steering," 2023, *arXiv:2311.06668*.

[28] H. Zhao et al., "MMICL: Empowering vision-language model with multi-modal in-context learning," in *Proc. Int. Conf. Learn. Representations*, 2024.

[29] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical domain hallucination test for large language models," in *Proc. 27th Conf. Comput. Natural Lang. Learn.*, 2023, pp. 314–334.

[30] J. Chen et al., "Detecting and evaluating medical hallucinations in large vision language models," 2024, *arXiv:2406.10185*.

[31] Z. Cao, Y. Yang, and H. Zhao, "SCANS: Mitigating the exaggerated safety for LLMs via safety-conscious activation steering," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 23523–23531.

[32] A. Zou et al., "Representation engineering: A top-down approach to AI transparency," 2023, *arXiv:2310.01405*.

[33] S. Liu, H. Ye, and J. Zou, "Reducing hallucinations in large vision-language models via latent space steering," in *Proc. Int. Conf. Learn. Representations*, 2025.

[34] I. Jolliffe, "Principal component analysis," *Springer Ser. Statist.*, vol. 2, 2002, Art. no. 100.

[35] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Informat. Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.

[36] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in *Proc. 2nd Workshop Shortcomings Vis. Lang.*, 2019, pp. 26–36.

[37] W. Chen et al., "Cross-modal causal intervention for medical report generation," *IEEE Trans. Image Process.*, vol. 34, pp. 2970–2985, 2025, doi: 10.1109/TIP.2025.3568746.

[38] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12910–12917.

[39] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13753–13762.

[40] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1439–1449.

[41] Z. Huang, X. Zhang, and S. Zhang, "KiUT: Knowledge-injected U-transformer for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19809–19818.

[42] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3334–3343.

[43] A. Liu, Y. Guo, J.-H. Yong, and F. Xu, "Multi-grained radiology report generation with sentence-level image-language contrastive learning," *IEEE Trans. Med. Imag.*, vol. 43, no. 7, pp. 2657–2669, Jul. 2024.

[44] Z. Lin et al., "Contrastive pre-training and linear interaction attention-based transformer for universal medical reports generation," *J. Biomed. Informat.*, vol. 138, 2023, Art. no. 104281.

[45] H. Xue, Q. Ma, G. Liu, J. Qu, Y. Liu, and A. Liu, "CLR2G: Cross modal contrastive learning on radiology report generation," in *Proc. 33 rd ACM Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 2742–2752.

[46] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024.

[47] M. Awais et al., "Foundation models defining a new era in vision: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2245–2264, Apr. 2025.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[49] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.
[50] R. Hendel, M. Geva, and A. Globerson, "In-context learning creates task vectors," *Findings Assoc. Comput. Linguistics: EMNLP*, pp. 9318–9333, Dec. 2023.
[51] E. Todd, M. L. Li, A. S. Sharma, A. Mueller, B. C. Wallace, and D. Bau, "Function vectors in large language models," 2023, *arXiv:2310.15213*.
[52] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
[53] F. Liu et al., "Aligning, autoencoding and prompting large language models for novel disease reporting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3332–3343, May 2025.
[54] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 3876–3887.
[55] H. Touvron et al., "LLAMA 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
[56] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.
[57] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 590–597.
[58] Q. Zhu, T. S. Mathai, P. Mukherjee, Y. Peng, R. M. Summers, and Z. Lu, "Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2023, pp. 189–198.
[59] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2002, pp. 311–318.
[60] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–457.
[61] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2004.
[62] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[63] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5904–5914.
[64] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11558–11567.
[65] K. Zhang et al., "A generalist vision–language foundation model for diverse biomedical tasks," *Nature Med.*, vol. 30, pp. 3129–3141, 2024.
[66] C. Peng, K. Zhang, M. Lyu, H. Liu, L. Sun, and Y. Wu, "Scaling up biomedical vision-language models: Fine-tuning, instruction tuning, and multi-modal learning," 2025, *arXiv:2505.17436*.
[67] K. Liu et al., "Enhanced contrastive learning with multi-view longitudinal data for chest X-ray report generation," 2025, *arXiv:2502.20056*.
[68] A. Sellergren et al., "Medgemma technical report," 2025, *arXiv:2507.05201*.

Mingjie Li received the PhD degree in computer science from the University of Technology Sydney. He is a Tenure-Track associate professor with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University (SJTU). Prior to joining SJTU, he was a postdoctoral research fellow with the Department of Psychiatry and Behavioral Sciences, Stanford University. His research interests include span machine learning, medical AI, and computer vision, with a specific focus on developing medical generative models and foundation models to empower precision oncology and neuroscience.

Rui Liu received the master's degree from Monash University with First Class Honors. She is currently working toward the PhD degree under the supervision of Prof. Xiaojun Chang with the University of Technology Sydney. She is a research assistant with Monash Children's Hospital. Her research interests include machine learning and artificial intelligence in public health and especially in medical image analysis.

Zeyi Shi received the master's degree from RMIT University, in 2022. She is currently working toward the PhD degree with the Faculty of Engineering and Information Technology, University of Technology Sydney. Her research interests include computer vision, multimodal learning, and generative AI.

Mingfei Han received the BEng degree from Nankai University, the MEng degree from the University of Chinese Academy of Sciences, and the PhD degree under the supervision of Prof. Xiaojun Chang from the University of Technology Sydney. He is a professor with Future Media Computing Lab, School of Information Science and Technology, University of Science and Technology of China. He was a postdoc scholar with the Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). His research interests include computer vision and machine learning, with a particular emphasis on large vision-language models and video object perception.

Lina Yao (Senior Member, IEEE) is a senior principal research scientist and science lead (research manager) with CSIRO's Data61, conjoint professor with UNSW, Honorary professor with Macquarie University and adjunct professor with the University of Technology Sydney. Her research interests include developing generalizable and explainable data-efficient data mining, machine learning and deep learning algorithms—as well as designing systems and interfaces—to enable novel ways of human-machine interactions

Zhihui Li received the PhD degree from the School of Computer Science and Engineering, University of New South Wales (UNSW), Sydney, NSW, Australia, in 2021. She is currently a professor with the Future Media Computing Lab, School of Information Science and Technology, University of Science and Technology of China. She has authored or co-authored more than 50 high-quality articles in top-tier venues, nine of which are identified as ESI highly cited papers. Her research interests include machine learning and its applications to computer vision and multimedia.

Xiaojun Chang (Senior Member, IEEE) is a professor with Future Media Computing Lab, School of Information Science and Technology, University of Science and Technology of China. He is also a visiting professor with the Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). Before joining USTC, he was a professor with Australian Artificial Intelligence Institute, University of Technology Sydney. After graduation, he subsequently worked as a postdoc research fellow with the School of Computer Science, Carnegie Mellon University, lecturer and senior lecturer with the Faculty of Information Technology, Monash University, Australia, and associate professor with the School of Computing Technologies, RMIT University. He has spent most of his time working on exploring multiple signals (visual, acoustic, textual) for automatic content analysis in unconstrained or surveillance videos. He has achieved top performances in various international competitions, such as TRECVID MED, TRECVID SIN, and TRECVID AVS.

Kilian M. Pohl received the PhD degree in computer science and artificial intelligence laboratory from the Massachusetts Institute of Technology. He is a professor in psychiatry & behavioral sciences and (by courtesy) in electrical engineering with Stanford University. He is also the founder of the Computational Neuroscience Laboratory (CNS^{Lab}), which focuses on creating machine learning technology to identify phenotypes improving personalized medicine and prevention of neuropsychiatric disorders. He is currently the principal investigator of several NIH funded projects and a senior editor of Medical Image Analysis. He was an instructor with Harvard Medical School, a research scientist with IBM Research, an assistant professor with the University of Pennsylvania, and a program director of Biomedical Computing at SRI International.

Md Tauhidul Islam received the PhD degree in electrical engineering from Texas A & M University, in 2018. He is an assistant professor with the Department of Radiation Oncology, Stanford University. His current research interests include biomedical omics data analysis using deep learning, manifold embedding, and the interpretability of deep neural networks.

Lei Xing received the PhD degree in physics from the Johns Hopkins University, Baltimore, MD, USA, in 1992. He completed the Medical Physics Training with the University of Chicago, Chicago, IL, USA. He is currently the Jacob Haimson professor and the director with the Medical Physics Division, Radiation Oncology Department, Stanford University, Stanford, CA, USA. He also holds affiliate faculty positions with the Department of Electrical Engineering, Bio-X and Molecular Imaging Program, Stanford University. He has been a member of the Radiation Oncology Faculty, Stanford University, since 1997. His current research interests include medical imaging, artificial intelligence in medicine, treatment planning, image guided interventions, nanomedicine, and applications of molecular imaging in radiation oncology.