

Article

Liver-VLM: Enhancing Focal Liver Lesion Classification with Self-Supervised Vision-Language Pretraining

Jian Song ^{1,2}, Yuchang Hu ¹, Hui Wang ^{3,*} and Yen-Wei Chen ^{2,*}

¹ School of Mathematical Sciences, Huaqiao University, Quanzhou 362021, China; songjian@hqu.edu.cn (J.S.); hu3191726921@gmail.com (Y.H.)

² College of Information Science and Engineering, Ritsumeikan University, Osaka 567-8570, Japan

³ School of Information Science and Engineering, Shandong University, Qingdao 266237, China

* Correspondence: wanghui0739@163.com (H.W.); chen@is.ritsumei.ac.jp (Y.-W.C.)

Abstract

Accurate classification of focal liver lesions (FLLs) is crucial for reliable clinical decision-making. Inspired by contrastive vision-language models such as CLIP and MedCLIP, we propose Liver-VLM for FLLs classification, trained on a dedicated multi-phase 2D CT dataset. Liver-VLM aligns multi-phase CT image representations with class-specific textual descriptions by calculating their similarity under a cross-entropy loss. Furthermore, we design tailored, enriched textual prompts to stabilize optimization and enable robust classification even with limited labeled data. Additionally, self-supervised pretraining and data augmentation strategies are incorporated to further improve classification performance. Experimental results on an in-house MPCT-FLLs dataset demonstrate that Liver-VLM consistently outperforms existing VLMs, achieving an accuracy of $85.63 \pm 3.18\%$ and an AUC of 0.94 ± 0.01 . Our findings highlight the efficacy of self-supervised learning and task-specific augmentation in overcoming data scarcity and distributional biases in medical image analysis.

Keywords: focal liver lesions (FLLs); multi-phase CT imaging; self-supervised learning; vision-language models (VLMs); data augmentation; phase shuffle strategy

Academic Editors: Marija Brkić Bakarić, Sanja Seljan and Ivan Dunder

Received: 20 May 2025

Revised: 25 October 2025

Accepted: 25 November 2025

Published: 27 November 2025

Citation: Song, J.; Hu, Y.; Wang, H.; Chen, Y.-W. Liver-VLM: Enhancing Focal Liver Lesion Classification with Self-Supervised Vision-Language Pretraining. *Appl. Sci.* **2025**, *15*, 12578. <https://doi.org/10.3390/app152312578>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Focal liver lesions (FLLs) are a common finding in multi-phase computed tomography (CT) images, ranging from benign conditions such as hepatic cysts to malignant tumors like hepatocellular carcinoma (HCC). Multi-phase CT images are widely used for the diagnosis of FLLs. Standard multi-phase CT [1] protocols typically acquire scans at three stages: Non-Contrast (NC), Arterial (ART), and Portal Venous (PV) phases, although some protocols may additionally include a Delayed (DL) phase. As depicted in Figure 1, several representative categories of FLLs, including Cyst, Focal Nodular Hyperplasia (FNH), hepatocellular carcinoma (HCC), and Hemangioma (HEM), are shown across the corresponding imaging phases. Notably, FLLs demonstrate phase-dependent visual variations due to differing contrast enhancement behaviors over time, highlighting the diagnostic importance of temporal phase information.

Accurate classification of FLLs from the multi-phase CT images is essential for effective clinical decision-making. However, the diagnostic process often depends heavily on radiologists' clinical experience, leading to inherent subjectivity. To address this

limitation, a variety of deep learning-based approaches [2,3] have been proposed and have achieved state-of-the-art performance. Despite these advances, the creation of annotated multi-phase CT datasets remains labor-intensive and time-consuming, resulting in limited data availability. Consequently, most medical imaging datasets are small-scale, posing a significant challenge for further improving model performance. In our previous work, we demonstrated that utilizing pretrained models on large-scale datasets such as ImageNet [4] or adopting self-supervised learning (SSL) techniques [5–7] can significantly enhance the classification accuracy of FLLs. It is worth noting that most existing methods rely on traditional fully connected (FC) layer-based classification.

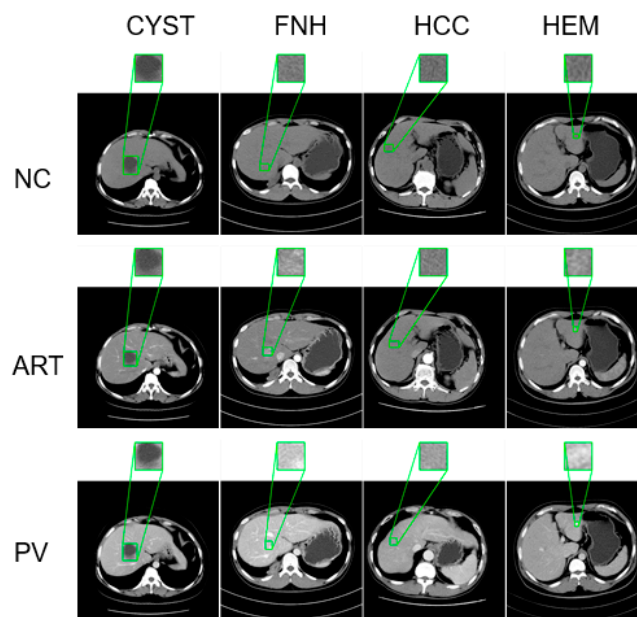


Figure 1. Evolutionary patterns of four FLL types across three phases.

In contrast, Contrastive Language–Image Pre-training (CLIP) [8] has been proposed to learn generalized visual representations from textual annotations, offering a paradigm known as text-guided image classification. In this approach, each class label is represented as a natural-language prompt, and classification is performed by computing the cosine similarity between the image embedding and each text embedding. The prompt with the highest similarity score is then selected as the predicted class. Compared to traditional FC-based methods, text-guided image classification offers two main advantages: (1) It allows the use of natural language descriptions (e.g., “a photo of a cat”, “an image of a dark dog”) to support fine-grained and context-aware classification, thereby enhancing prediction accuracy. (2) Trained on a large and diverse set of imaging–text pairs, Vision-Language Models (VLMs) acquire broad semantic knowledge, resulting in superior generalization across different tasks and domains.

However, CLIP’s performance on medical image classification remains suboptimal due to the domain gap between natural and medical images. MedCLIP [9] extends CLIP to the medical domain through a text-guided supervision strategy, where a semantic similarity matrix derived from medical reports serves as a teacher signal. The model learns to align its predicted image–text similarity matrix with the semantic matrix to encode medical semantics more effectively. Nevertheless, being pretrained mainly on X-ray images, MedCLIP faces domain shift issues when generalized to CT scans, and its more expressive design further increases computational cost and optimization complexity. Beyond CLIP-style models, recent attention-based frameworks have also

advanced liver CT analysis. For example, Attention GhostUNet++ [10] employs multi-scale channel-spatial-depth attention to improve liver segmentation, and Uni-MLIP [11] leverages unified self-supervised vision-language pretraining with efficient attention to refine cross-modal representations. These works highlight the growing trend of attention-driven designs while emphasizing the need for more lightweight attention strategies and domain-adaptive solutions suitable for liver-specific representation learning.

To address these challenges, we propose Liver-VLM for FLLs classification. While inspired by the CLIP framework, Liver-VLM is specifically trained on a dedicated FLLs dataset, making it better suited for the medical imaging domain. The primary novelty of Liver-VLM does not lie in proposing a new, complex VLM architecture, but in demonstrating how a strategically designed yet simple VLM framework can be effectively tailored and validated for the specific, clinically critical task of FLLs classification.

The main contributions are summarized below:

- (1) We propose Liver-VLM, a Vision-Language Model for FLLs classification, which is specifically trained on a dedicated multi-phase CT FLLs dataset, making it better suited for the multi-phase CT domain. We focus on clinically problem-driven design over architectural complexity and position Liver-VLM as a strong, reproducible, and clinically interpretable baseline model for liver lesion diagnosis.
- (2) We design a self-supervised pre-training strategy tailored for Liver-VLM by integrating the phase shuffle prediction task, which was proposed in our previous work [6]. Our work is the first to integrate this specific three-channel phase-shuffle pretext task directly into a VLM training pipeline for medical imaging. Previous works used it for pre-training a visual-only model. We demonstrate that this task, which is inherently aligned with the multi-phase nature of Live CT, is highly effective not only for learning visual representations but for aligning those representations with clinical language, thereby enhancing the final VLM's discriminative power.
- (3) We design tailored enriched textual prompts to stabilize optimization and enable robust classification even with limited labeled data. We also propose a data augmentation technique based on phase shuffle, which expands the training dataset by generating all six possible permutations of the CT phase order. This method enhances data diversity and improves the model's robustness to variations in phase presentation.

Preliminary work was accepted as a conference paper at the 2025 International Conference on Innovation in Medicine and Healthcare and has also been presented on arXiv [12]. This paper involves methodological and experimental extensions and validations beyond the initial work. Specifically, we design a self-supervised pre-training strategy tailored for Liver-VLM by introducing a phase-shuffle prediction task (contribution 2) and propose a data augmentation technique based on phase shuffle (contribution 3). Both contributions lead to significant improvements in classification accuracy.

This paper is structured as follows: Section 2 provides a brief overview of related work. In Section 3, we present a detailed description of the proposed approach. The experimental setup and results are presented in Section 4, and the conclusions are presented in the final section.

2. Related Work

2.1. CLIP

CLIP [8] is a model introduced by OpenAI that leverages contrastive learning to align images and texts in a shared embedding space. The architecture of CLIP consists of an image encoder (typically using ResNet50 [13] or ViT [14]) and a text encoder (usually

based on the Transformer structure [15–17]). Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings are the correct matches. During training, CLIP aims to maximize the cosine similarity between embeddings of the N positive image-text pairs, while reducing the similarity among the $N^2 - N$ negative pairs through contrastive learning. Specifically, the model optimizes a symmetric InfoNCE loss, encouraging the embeddings of matched image-text pairs to be close in the multimodal space while pushing apart the mismatched pairs. It is trained on large-scale image-text pairs to learn a universal cross-modal representation, thereby minimizing reliance on task-specific data and enabling rapid adaptation to diverse tasks without requiring labeled data. This representation also enables the model to perform various vision tasks, such as image classification, retrieval, and generation. Despite its strengths, CLIP's performance on medical image classification remains suboptimal due to the domain gap between natural and medical images.

2.2. MedCLIP

MedCLIP [9] is a contrastive vision-language pretraining framework specifically designed for the medical domain. Its core innovation lies in constructing a semantic similarity matrix based on medical knowledge graphs (such as UMLS [18]), which encodes soft relational labels among diseases, symptoms, and findings. During training, MedCLIP aligns the predicted image-text similarity distribution with this semantic matrix via a cross-entropy (CE) loss, thereby encouraging the model to capture clinically meaningful cross-modal relationships. To process grayscale medical images, MedCLIP replicates single-channel inputs into pseudo-RGB format, ensuring compatibility with standard backbones such as Vision Transformer (ViT) and ResNet50 pretrained on ImageNet. On the textual side, Bio_ClinicalBERT [17] or similar domain-adapted models are used to enhance understanding of medical terminology and report semantics.

While this design effectively transfers CLIP's contrastive paradigm into the medical domain, it introduces additional supervision through a semantic similarity matrix, leading to higher computational cost and more complex optimization.

3. Method

3.1. Overview of the Proposed Method

In contrast to large-scale CLIP-style models trained on massive image-text pairs and other attention-heavy or adapter-based VLMs, we propose Liver-VLM, a lightweight and domain-specific framework tailored for multi-phase CT liver analysis. By leveraging pretrained representations, the model minimizes architectural overhead and supports efficient fine-tuning on limited medical data. The overview of Liver-VLM is shown in Figure 2. Similarly to CLIP or other VLMs, the model consists of a ResNet-18 image encoder and a pre-trained BERT-Base uncased [19] text encoder, where only BERT's lightweight multi-head attention is used on the textual side without explicit visual attention.

In our preliminary work [12], we demonstrated that initializing the image encoder with ImageNet pretraining improved classification performance. To further enhance accuracy and robustness, we propose a self-supervised pretraining strategy tailored for Liver-VLM, which introduces a phase shuffle prediction task (Figure 2a). This task requires the model to predict the correct order of shuffled CT phases, enabling the encoder to learn phase-aware temporal representations across NC, ART, and PV phases.

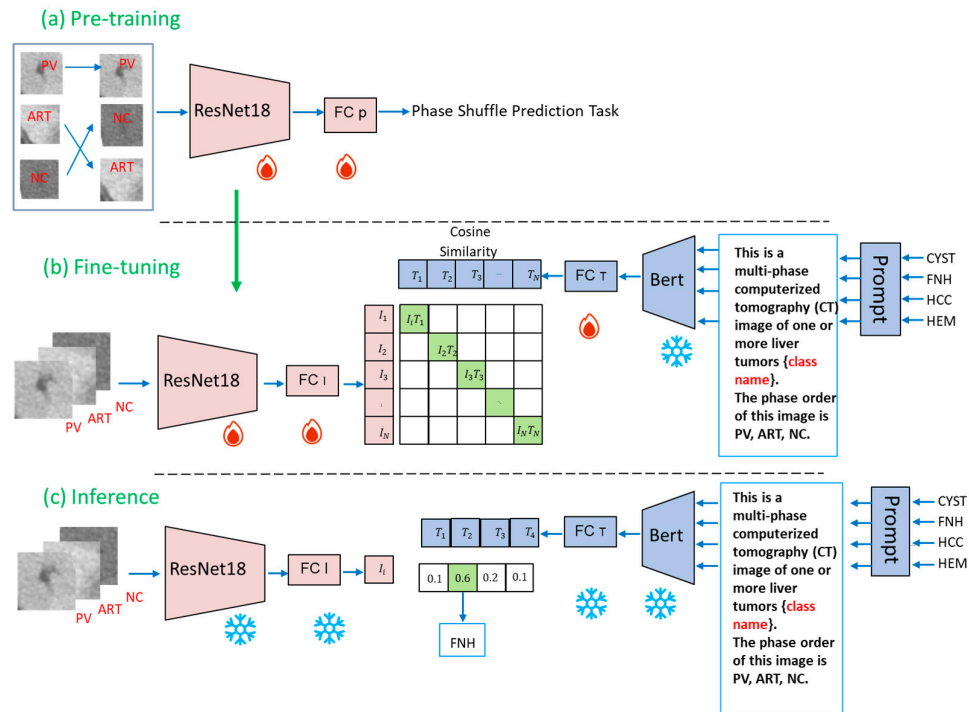


Figure 2. Overview of Liver-VLM, (a) Pre-training (b) Fine-tuning (c) Inference. (CYST: Cyst; FNH: Focal Nodular Hyperplasia; HCC: Hepatocellular Carcinoma; HEM: Hemangioma; PV: Portal Venous; ART: Arterial; NC: Non-Contrast).

During target training (fine-tuning), as illustrated in Figure 2b, cosine similarity is computed between image features and class-specific text embeddings, and a CE loss is applied to maximize similarity for correct image-text pairs while minimizing it for incorrect ones, thereby enabling effective multimodal alignment. To further address the challenge of semantic similarity in CLIP, particularly in scenarios with a limited number of lesion categories, we design enriched prompts that incorporate complete class names, imaging phase information, and domain-specific medical terminology. These enriched prompts enhance the semantic expressiveness of the text input and improve the model's ability to align visual and textual representations. Furthermore, a phase-shuffling-based data augmentation strategy is introduced to increase data diversity and strengthen model robustness by generating all six permutations of the CT phase order.

In the inference stage (Figure 2c), a test image (FLL) is classified into the category with the highest similarity to the corresponding textual description.

3.2. Self-Supervised Pre-Training with Phase Shuffle Prediction Task

Transfer learning allows models to adapt knowledge from one task to another, accelerating convergence and improving performance with limited annotated data. However, due to domain differences, ImageNet-pretrained models may be suboptimal for medical imaging tasks. To address this, SSL pretrains models directly on target-domain data using pretext tasks, enabling better feature representation without requiring annotations. The model is then fine-tuned on the target task with limited labeled data. Building on our prior research, where we demonstrated the effectiveness of phase-shuffle prediction for capturing meaningful temporal features from multi-phase CT scans [6,7], we apply this tailored pretext task for SSL in this work. As shown in Figure 2a, the phase order of multi-phase CT images is randomly shuffled, and the model is tasked with predicting the correct phase order. For instance, the original phase order is PV, ART, and

NC, but in Figure 2a, it is shuffled to PV, NC, and ART. Since there are three phases, there are 3! (i.e., six) possible permutations, leading to a six-class classification problem, with each class corresponding to a specific phase order (Class 1: PV, ART, NC; Class 2: PV, NC, ART; Class 3: ART, NC, PV; Class 4: ART, PV, NC; Class 5: NC, ART, PV; Class 6: NC, PV, ART).

To perform the prediction, the shuffled images of three-phase CT images are treated as a three-channel input, similar to a color image, and fed into the image encoder (ResNet) for feature extraction. The input size is $3 \times 224 \times 224$, and the encoder outputs a feature vector of size 512×1 . An FC layer denoted as FC_p is then applied to predict the shuffled phase order with the output layer consisting of six neurons corresponding to the six possible classes.

The model is trained using the CE loss, a standard loss function for classification tasks, which measures the discrepancy between the predicted probability distribution and the true class labels. The CE loss is formally defined in Equation (1):

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^M \sum_{c=1}^{C1} y_{i,c} \log(p_{i,c}) \quad (1)$$

where M is the number of samples used for pretraining, and $C1 = 6$ is the number of possible permutations of the three CT phasea, and $y_{i,c}$ is the true label indicating whether the i th sample belongs to the class c ; if it does, then $y_{i,c} = 1$, otherwise $y_{i,c} = 0$. $p_{i,c}$ denotes the predicted probability that the i th sample belongs to the class c .

3.3. Target Training (Fine-Tuning)

After the self-supervised pre-training phase, the model is fine-tuned for the target task of FLLs classification, as illustrated in Figure 2b. The pre-trained weights of the image encoder (ResNet) are employed for initialization and further optimized with a limited number of labeled multi-phase CT images. This fine-tuning process effectively transfers the pre-trained representations to the downstream classification task, thereby improving accuracy and robustness while mitigating overfitting due to limited annotations. The visual embeddings extracted by the image encoder are then projected through an FC layer (FC_l) to match the dimensionality of the text embeddings, which allows dot-product similarity computation between modalities.

To further enhance the semantic correspondence between images and textual labels, textual prompts are incorporated during fine-tuning. Since short or context-deficient labels may lead to semantic ambiguity, prompt engineering strategies inspired by CLIP are adopted. In the original CLIP framework, simple prompts such as “This is a photo of a dog” were commonly used. To provide richer contextual information in the medical domain, abbreviated lesion and phase names (e.g., ‘FNH’, ‘HCC’, ‘HEM’, ‘PV’, ‘ART’, ‘NC’) are replaced with their corresponding full medical terms (e.g., ‘Focal Nodular Hyperplasia’, ‘Hepatocellular Carcinoma’, ‘Hemangioma’, ‘Portal Venous’, ‘Arterial’, ‘Non-Contrast’). Accordingly, a more informative prompt is designed as follows: “This is a multi-phase computerized tomography (CT) image of one or more liver tumors {class name}, where the phase order of this image is arterial phase, portal venous phase, non-contrast phase.”

The designed prompts are encoded using a BERT-based uncased text encoder to obtain text embeddings. To mitigate overfitting caused by limited annotations in medical datasets, the text encoder is kept frozen during training, and only the FC layer is appended to BERT (FC_t) is updated.

By aligning the projected image and text embeddings in a shared latent space, the model learns to associate visual patterns with their corresponding semantic descriptions, thereby improving its ability to discriminate between different lesion categories. Considering the small dataset size, limited number of lesion classes, and subtle inter-class

semantic variations, a standard CE loss is adopted instead of the symmetric contrastive loss employed in the original CLIP framework. The CE loss is defined as shown in Equation (2):

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(s_{i,c}) \quad (2)$$

$$s_{i,c} = \frac{\mathbf{I}_i \cdot \mathbf{T}_c}{\|\mathbf{I}_i\| \|\mathbf{T}_c\|} \quad (3)$$

where N is the number of samples used for target training (fine-tuning), and $C = 4$ is the number of possible categories of FLLs, and $y_{i,c}$ is the true label indicating whether the i th sample belongs to the class c ; if it does, then $y_{i,c} = 1$, otherwise $y_{i,c} = 0$. $s_{i,c}$ denotes the predicted similarity between the i th sample and the textual description of the c th category, calculated using Equation (3). $\mathbf{I}_i$ and $\mathbf{T}_c$ denote the feature vectors of the i th image and the c th textual description, extracted by ResNet18 and BERT, respectively. $\|\mathbf{I}_i\|$ and $\|\mathbf{T}_c\|$ denote the L2 norms of $\mathbf{I}_i$ and $\mathbf{T}_c$.

After fine-tuning, the encoder is kept frozen during inference to maintain consistent and stable feature extraction, as shown in Figure 2c.

3.4. Data Augmentation Based on Phase Shuffle

Building upon the enriched textual prompts with imaging phase information described in Section 3.3, we propose a data augmentation strategy based on phase shuffling. This approach aims to increase data diversity and enhance model robustness by generating all possible permutations of the CT phase order. Specifically, for three imaging phases, there are $3! = 6$ possible permutations, resulting in six augmented versions of each original image. As shown in Figure 3, each augmented image is paired with a textual prompt that reflects the corresponding phase sequence. The six permutations are as follows: Data 1 (original): NC, ART, PV; Data 2: NC, PV, ART; Data 3: PV, ART, NC; Data 4: PV, NC, ART; Data 5: ART, NC, PV; Data 6: ART, PV, NC.

Each augmented sample preserves the image content while varying the phase order in the associated prompt. This design allows the model to learn phase-invariant representations, thereby improving generalization across different CT phase-sequences.

	Image Encoder			Text Encoder
	Channel 1	Channel 2	Channel 3	Phase order in Prompt
Original Data augmentation				{NC, ART, PV}
				{NC, PV, ART}
				{PV, ART, NC}
				{PV, NC, ART}
				{ART, NC, PV}
				{ART, PV, NC}

Figure 3. Data augmentation strategies and corresponding prompt pairs.

4. Experimental Results

4.1. Dataset and Implementations

We assess the performance of the proposed Liver-VLM using an internally curated Multi-Phase CT dataset of Focal Liver Lesions (MPCT-FLLs) [20]. The dataset comprises four lesion types: Cyst, FNH, HCC, and HEM, collected from Sir Run Run Shaw Hospital,

affiliated with Zhejiang University, during the period of 2015 to 2017. In total, 85 multi-phase CT volumes were collected, from which 489 representative 2D slice images centered on the lesion regions were manually selected. The slice thickness is 5 or 7 mm, and the in-plane resolution of 0.57–0.59 mm. Each axial 2D slice has a spatial resolution of 512×512 pixels and includes three distinct contrast phases: NC, ART, and PV. The regions of interest (ROIs) corresponding to the lesions were delineated by experienced radiologists. For model training and inference, the extracted 2D ROIs were resized to 224×224 pixels. The three-phase 2D CT images were then organized into three-channel composite inputs, resulting in input tensors with dimensions of $3 \times 224 \times 224$. Table 1 presents the data distribution across five groups used for five-fold cross-validation in model evaluation. In each fold, one group served as the test set, while the remaining four were used for training. Accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were adopted as the evaluation metrics. The final performance metrics were obtained by averaging the results across all five folds.

For fine-tuning, we train our network for 200 epochs with a batch size of 32, using the Adam [21] optimizer and a learning rate of 0.0005. The training setup is summarized in Table 2.

Table 1. Dataset distribution for five-fold cross-validation.

		CYST	FNH	HCC	HEM	Total
G1:	case	5	4	4	4	17
	slice	29	15	30	21	95
G2:	case	6	3	4	4	17
	slice	31	17	29	33	110
G3:	case	6	3	4	4	17
	slice	37	7	36	17	97
G4:	case	6	3	4	4	17
	slice	24	17	35	19	95
G5:	case	7	3	3	4	17
	slice	28	20	32	12	92
Total:	case	30	16	19	20	85
	slice	149	76	162	102	489

Table 2. Computation environment.

GPU	NVIDIA RTX A6000 (NVIDIA, Santa Clara, CA, USA)
CPU	Intel(R) Core(TM) i9-10980XE @ 3.00 GHz (Intel, Santa Clara, CA, USA)
OS	Ubuntu 20.04.5 LTS
Deep Learning Framework	Pytorch 2.1.1

4.2. Results

4.2.1. Ablation Studies

Ablation experiments were conducted on the MPCT-FLLs dataset to evaluate the effectiveness of each component in the proposed model. Model performance was assessed under the same evaluation protocol described in Section 4.1.

Table 3 presents the ablation results comparing different pre-training strategies. When the original CLIP model is directly applied to our dataset in a zero-shot setting using a single medical prompt, it achieves an average accuracy of $20.54 \pm 5.61\%$ and an AUC of 0.43 ± 0.09 , indicating suboptimal performance. Using an ensemble of 16 medical

prompts, as per standard practice, CLIP achieves an average accuracy of $18.74 \pm 2.87\%$ and an AUC of 0.42 ± 0.07 . This slight degradation suggests that simply aggregating multiple generic medical prompts cannot effectively mitigate the domain shift between natural-image pretraining and multi-phase CT imaging. In contrast, all three variants of the proposed Liver-VLM framework, trained on the MPCT-FLLs dataset, demonstrate substantial improvements in classification performance. Training the model from scratch yields an accuracy of $81.15 \pm 7.36\%$ and an AUC of 0.92 ± 0.04 . When using ImageNet pre-training for the ResNet18 image encoder, accuracy and AUC improve to $83.35 \pm 4.81\%$ and 0.92 ± 0.03 , respectively. With the proposed self-supervised pre-training strategy, in which the image encoder is trained from scratch using only self-supervised learning without any ImageNet initialization, the model further achieves an accuracy of $85.09 \pm 3.89\%$ and an AUC of 0.92 ± 0.03 . These results clearly demonstrate the effectiveness of pre-training the image encoder, particularly with the proposed self-supervised learning method.

Table 3. Ablation results comparing different pre-training strategies.

Model	Avg. Acc (%)	AUC
CLIP (zero-shot inference) [8] single medical prompt	20.54 ± 5.61	0.43 ± 0.09
CLIP (zero-shot inference) [8] ensemble of 16 medical prompts	18.74 ± 2.87	0.42 ± 0.07
Live-VLM (from scratch) [12]	81.15 ± 7.36	0.92 ± 0.04
Live-VLM (ImageNet) [12]	83.35 ± 4.81	0.92 ± 0.03
Liver-VLM (Self-Supervised)	85.09 ± 3.89	0.92 ± 0.03

Bold indicates the highest values.

To rule out the possibility that the performance gain stems solely from an increased number of pre-training samples, we performed a control experiment using SimSiam [22] with the same total optimization steps as our phase-shuffle SSL. The results are summarized in Table 4. As shown, when both encoders were fine-tuned under identical VLM settings, our phase-shuffle model achieved an accuracy of $85.09 \pm 3.89\%$, compared to $80.19 \pm 5.16\%$ for SimSiam. This confirms that the performance improvement originates from the phase-order signal rather than additional optimization steps. The comparable AUCs (both around 0.92 ± 0.03) indicate similar overall discriminative capability, whereas the higher accuracy of the phase-shuffle model reflects more consistent decision boundaries, suggesting better phase-aware feature alignment learned during SSL.

Table 4. Control experiment of self-supervised Liver-VLM under the proposed prompt.

Model	CYST	FNH	HCC	HEM	Avg. Acc (%)	AUC
Liver-VLM (pre-train using SimSiam SSL)	96.50 ± 5.54	92.79 ± 6.04	71.05 ± 11.23	58.46 ± 20.96	80.19 ± 5.16	0.92 ± 0.03
Liver-VLM (proposed method) (pre-train using phase shuffle SSL)	95.79 ± 3.54	87.60 ± 9.54	80.99 ± 10.67	68.54 ± 24.83	85.09 ± 3.89	0.92 ± 0.03

Bold indicates the highest values.

To further assess the competitiveness of the ResNet18 [13] image encoder used in our Liver-VLM, we additionally compared it with several recent hybrid architectures, including Video-Swin Transformer [23], ConvNeXt-Tiny [24], and EfficientNet-V2-S [25], which have shown strong performance in natural and medical image analysis. As shown in Table 5, the proposed ResNet-18 model achieved $85.09 \pm 3.89\%$ accuracy and 0.92 ± 0.03 AUC, which is comparable to or even slightly higher than EfficientNet-V2-S (84.52 ± 2.39

%, 0.93 ± 0.03) and clearly surpasses ConvNeXt-Tiny (70.26 ± 9.35 %, 0.87 ± 0.04) and Video-Swin (62.81 ± 14.42 %, 0.82 ± 0.08). These results confirm that despite its lightweight nature, the ResNet-18 backbone remains highly competitive for multi-phase CT liver lesion classification.

Table 5. Ablation results comparing the Image Encoder of the self-supervised Liver-VLM under the proposed prompt.

Image Encoder	CYST	FNH	HCC	HEM	Avg. Acc (%)	AUC
Liver-VLM (Video_Swin)	97.09 ± 2.70	57.63 ± 39.40	47.95 ± 25.09	32.94 ± 22.12	62.81 ± 14.42	0.82 ± 0.08
Live-VLM (ConvNeXt)	96.48 ± 3.94	75.13 ± 27.52	57.23 ± 20.64	40.36 ± 33.83	70.26 ± 9.35	0.87 ± 0.04
Liver-VLM(EfficientNet)	97.86 ± 4.28	80.69 ± 21.75	84.44 ± 13.85	57.52 ± 23.24	84.52 ± 2.39	0.93 ± 0.03
Liver-VLM (ResNet18)	95.79 ± 3.54	87.60 ± 9.54	80.99 ± 10.67	68.54 ± 24.83	85.09 ± 3.89	0.92 ± 0.03

Bold indicates the highest values.

To further evaluate the robustness of the proposed Liver-VLM under realistic conditions, we tested its performance on a misregistered dataset in which the three CT phases are inherently misaligned due to liver motion and variations in z-position. As shown in Table 6, the model achieved an accuracy of $84.98 \pm 5.66\%$ and an AUC of 0.91 ± 0.04 , which are comparable to the results on the perfectly aligned dataset ($85.09 \pm 3.89\%$, 0.92 ± 0.03). This indicates that the model remains stable and effective even when phase alignment is imperfect, highlighting its robustness and potential applicability in real clinical settings.

Table 6. Ablation study of Liver-VLM on aligned versus naturally misregistered ROIs under the proposed prompt.

Model	CYST	FNH	HCC	HEM	Avg. Acc (%)	AUC
Liver-VLM (Self-Supervised) (misregistered)	94.81 ± 4.41	72.39 ± 12.25	90.66 ± 6.59	67.88 ± 25.77	84.98 ± 5.66	0.91 ± 0.04
Liver-VLM (Self-Supervised)	95.79 ± 3.54	87.60 ± 9.54	80.99 ± 10.67	68.54 ± 24.83	85.09 ± 3.89	0.92 ± 0.03

Bold indicates the highest values.

Table 7 presents the ablation results evaluating the impact of our enriched prompt and data augmentation. Under the proposed self-supervised Liver-VLM configuration, we conducted three experiments: (1) Model 0: using a simple CLIP-style text prompt (i.e., “A CT scan of tumors {class name}.”); (2) Model 1: using our enriched text prompt as described in Section 3.3; (3) Model 2: applying data augmentation based on phase shuffling (the proposed method). As shown in Table 7, when using the simple CLIP-style prompt (Model 0), the model achieves an average accuracy of $82.46 \pm 7.34\%$ and an AUC of 0.93 ± 0.03 . With the enriched text prompt (Model 1), performance improves to $85.09 \pm 3.89\%$ accuracy and an AUC of 0.92 ± 0.03 . When data augmentation is also applied (Model 2), the model further improves, achieving $85.63 \pm 3.18\%$ accuracy and an AUC of 0.94 ± 0.01 . These results clearly demonstrate the effectiveness of both the enriched text prompt and the proposed data augmentation technique.

Table 7. Ablation of enriched prompt and data augmentation.

Model	CYST	FNH	HCC	HEM	Avg. Acc (%)	AUC
Model 0 Liver-VLM (CLIP Prompt)	92.97 ± 8.27	85.96 ± 4.51	73.7 ± 5.58	64.95 ± 11.12	82.46 ± 7.34	0.93 ± 0.03
Model 1	95.79 ± 3.54	87.60 ± 9.54	80.99 ± 10.67	68.54 ± 24.83	85.09 ± 3.89	0.92 ± 0.03

Liver-VLM (Proposed Prompt)

Model 2 (proposed method)

Liver-VLM (Proposed prompt + Data Augmentation)	98.03 ± 2.79	83.22 ± 21.07	83.45 ± 15.7	72.74 ± 10.51	85.63 ± 3.18	0.94 ± 0.01
---	---------------------	---------------	---------------------	----------------------	---------------------	--------------------

Bold indicates the highest values.

To assess the effect of data augmentation in the proposed self-supervised Liver-VLM with the proposed prompt, fold-level paired *t*-tests [26], and sample-level DeLong tests [27] were performed. Model 1 (without augmentation) yielded significantly higher AUCs for the CYST ($p_t = 0.0008$, $p_{DeLong} \approx 0$) and FNH ($p_t = 0.038$, $p_{DeLong} = 0.011$) classes, whereas Model 2 (with augmentation) achieved superior performance for HCC ($p_t < 0.001$, $p_{DeLong} = 0.017$) and slightly better results for HEM ($p_t < 0.001$, $p_{DeLong} = 0.06$). These findings suggest that the proposed augmentation enhances discrimination of malignant lesions (HCC) but may not provide consistent benefits for benign lesion classification.

To investigate whether the observed performance gains from phase-shuffle augmentation are specific to cross-modal learning, we trained a CNN-only classifier (ResNet18 + FC) using the augmented dataset. As shown in Table 8, the CNN-only classifier achieved $80.1 \pm 3.58\%$ accuracy and 0.93 ± 0.02 AUC, which are notably lower than the Liver-VLM using the same augmented data with enriched prompts ($85.63 \pm 3.18\%$ accuracy, 0.94 ± 0.01 AUC). This confirms that phase-shuffle augmentation primarily benefits cross-modal alignment in the VLM setting, rather than providing significant improvements for single-modality image classification.

Table 8. Comparison of Liver-VLM and CNN-only classifier performance with phase-shuffle augmentation.

Model	CYST	FNH	HCC	HEM	Avg. Acc (%)	AUC
Data augmentation with FC layer	98.49 ± 1.97	77.32 ± 13.46	83.04 ± 7.90	48.60 ± 15.93	80.1 ± 3.58	0.93 ± 0.02
Liver-VLM (proposed method) (Proposed prompt + Data Augmentation)	98.03 ± 2.79	83.22 ± 21.07	83.45 ± 15.7	72.74 ± 10.51	85.63 ± 3.18	0.94 ± 0.01

Bold indicates the highest values.

To verify whether the regions attended by the model correspond to clinically relevant lesion features, we applied Grad-CAM [28] to all four lesion types using a consistent color scale (cool colors indicate low contribution, warm colors indicate high contribution), as shown in Figure 4. The network consistently focuses on lesion margins and internal heterogeneity rather than the overall liver shape or surrounding vessels. Specifically, CYST shows a continuous ring of high response along the cyst wall with suppressed signal in the cavity; FNH peaks at the central scar and radiating strands; HCC highlights irregular peripheral rings and heterogeneous internal patches; HEM exhibits high peripheral nodules with centripetal filling. These patterns provide both intra-class interpretability and clear inter-class distinctions, confirming that Liver-VLM captures clinically relevant lesion features.

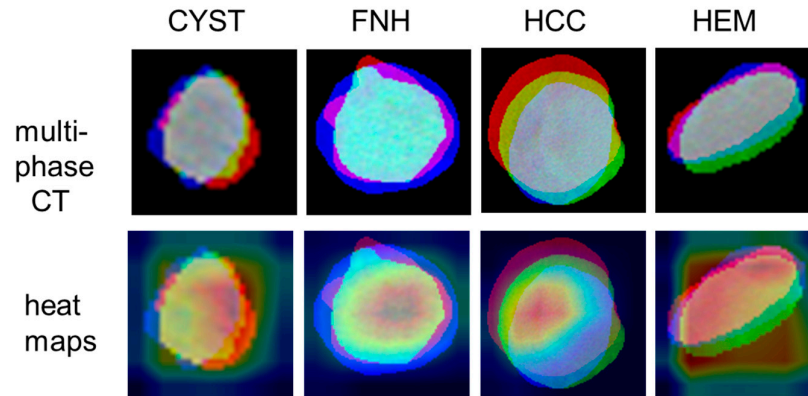


Figure 4. Grad-CAM visualization of Liver-VLM attention on four liver lesion types. Red denotes notes.

The image encoder, pretrained with Phase Shuffle SSL, is fine-tuned alongside the VLM during 4-class training using cross-entropy loss between image-text similarity and ground-truth labels. Long-text prompts and data augmentation are applied to improve robustness. Figure 5 presents pseudo-code illustrating the overall training and fine-tuning workflow.

```

1 # I_enc, T_enc: image and text encoders
2 # τ: temperature parameter
3 # Augment(.): data augmentation
4 # CrossEntropyLoss(.): classification loss
5
6 for k = 1, ..., K do                                # training epochs
7     I_aug = Augment(X_T)                            # apply data augmentation
8     I_feat = I_enc(I_aug)                          # extract image features
9     T_feat = T_enc(T)                              # extract text features
10    logits = I_feat · T_featT × exp(τ)             # compute similarity logits
11    L = CrossEntropyLoss(logits, y_T)                # compute loss
12    update(I_enc, T_enc, L)                          # update model parameters
13 end for
14
15 # inference
16 C_feat = T_enc(T_cls)                             # encode class text prompts
17 for each image X_E in test set do
18     I_feat = I_enc(X_E)                            # extract image features
19     sims = I_feat · C_featT × exp(τ)             # compute similarities
20     probs = softmax(sims)                          # obtain class probabilities
21     pred = argmax(probs)                           # predicted label
22 end for
23
24 return pred                                         # model prediction

```

Figure 5. Pseudo-code of the proposed self-supervised Liver-VLM.

4.2.2. Comparison with Other Models

We also compared the proposed method with several existing approaches, including zero-shot learning methods such as CLIP [8], MedCLIP [9], as well as our previously proposed models: Liver-VLM trained from scratch [12], Liver-VLM with ImageNet pretraining [12], and SSL using phase shuffle prediction with a traditional FC layer [7]. The comparison results are summarized in Table 9. As shown, applying the zero-shot learning models (i.e., CLIP and MedCLIP) directly to our dataset yielded relatively poor performance, with average accuracies of $20.54 \pm 5.61\%$ and $26.12 \pm 2.28\%$, and corresponding AUCs of 0.43 ± 0.09 and 0.48 ± 3.45 , respectively. In contrast, all models trained on the MPCT-FLL dataset significantly outperformed the zero-shot methods. In our prior work, we observed that ResNet18 yielded better performance compared to ResNet50, which was used in CLIP and MedCLIP. Among all evaluated methods, the proposed self-supervised Liver-VLM achieved the highest performance, with an accuracy of $85.63 \pm 3.18\%$ and an AUC of 0.94 ± 0.01 .

Table 9. Comparison with existing approaches.

Model		CYST	FNH	HCC	HEM	Acc (%)	AUC
CLIP (Zero-shot inference) [8]		0.00	0.00	0.00	100.00	20.54 ± 5.61	0.43 ± 0.09
MedCLIP (Zero-shot inference) [9]		31.94 ± 3.60	0.00	0.00	15.63 ± 5.58	26.12 ± 2.28	0.48 ± 3.45
Liver-VLM (from scratch) [12]	ResNet50	96.50 ± 5.54	90.79 ± 7.06	76.49 ± 13.52	39.08 ± 22.72	77.64 ± 7.83	0.89 ± 0.04
	ResNet18	97.32 ± 4.14	90.44 ± 8.95	76.02 ± 16.46	55.14 ± 28.83	81.15 ± 7.36	0.92 ± 0.04
Liver-VLM (ImageNet) [12]	ResNet50	95.94 ± 4.17	68.01 ± 30.87	78.11 ± 13.55	76.30 ± 16.96	83.05 ± 4.44	0.93 ± 0.02
	ResNet18	95.94 ± 4.17	73.72 ± 25.15	76.44 ± 12.60	70.41 ± 16.90	83.35 ± 4.81	0.92 ± 0.03
Phase shuffle prediction with a traditional FC layer [7]		98.27 ± 1.42	86.22 ± 12.27	82.90 ± 5.84	63.72 ± 12.68	84.82 ± 1.99	0.93 ± 0.02
Liver-VLM (ResNet18) (proposed)		98.03 ± 2.79	83.22 ± 21.07	83.45 ± 15.7	72.74 ± 10.51	85.63 ± 3.18	0.94 ± 0.01

Bold indicates the highest values.

5. Conclusions

In this study, we proposed Liver-VLM, an enhanced vision-language framework tailored for the classification of focal liver lesions in multi-phase CT images. Liver-VLM integrates a self-supervised pretraining strategy and task-specific data augmentation to better address the challenges of limited annotations and domain discrepancies in medical imaging. By adopting CE loss and enriched text prompts, the model achieves more stable training and enhanced alignment between CT images and clinical descriptions. Extensive experiments on an in-house MPCT-FLL dataset demonstrate the superiority of Liver-VLM over existing VLMs, with notable improvements in both accuracy and AUC. These results validate the effectiveness of combining self-supervised learning and augmentation in improving model generalizability for medical image understanding.

Our phase-shuffle data augmentation is designed for standard triphasic CT (arterial, portal venous, delayed), where the three phases can be permuted to generate diverse training samples. In clinical practice, however, CT protocols may vary: some scans include additional delayed or dual-arterial phases, while others may omit certain standard phases. Such variable-length multi-phase series can be accommodated by shuffling the available phases or subsequences, and the prompt template can be dynamically adapted to reflect the actual phase order (e.g., “A CT scan with phases {phase1}, {phase2}, ... of tumor {class}”). This flexibility allows phase-shuffle augmentation to generalize beyond standard triphasic protocols, although developing self-supervised methods applicable to all medical images remains a future direction.

Compared with traditional liver lesion classification methods (Table 9, reference [7]), Liver-VLM achieves a notable improvement in both accuracy and AUC, demonstrating its potential to provide more reliable decision support in clinical workflows. By leveraging multi-phase CT data and class-specific textual prompts, the proposed framework can assist radiologists in improving diagnostic consistency, reducing manual annotation effort, and mitigating biases due to limited or heterogeneous datasets.

Nevertheless, the model may fail in cases with very small lesions, low contrast between the lesion and liver parenchyma, or non-standard scanning protocols, which highlights the importance of careful clinical validation and further model refinement.

In the future, we plan to collect data from other institutions to serve as an external test cohort for further validation of the model’s generalizability.

Author Contributions: Conceptualization, Y.-W.C.; methodology, H.W. and J.S.; software, J.S. and Y.H.; validation, J.S., H.W., and Y.H.; formal analysis, J.S. and Y.-W.C.; investigation, J.S. and H.W.; resources, J.S.; data curation, Y.-W.C. and J.S.; writing—original draft preparation, J.S.; writing—review and editing, Y.-W.C. and J.S.; visualization, J.S., H.W., and Y.H.; supervision, Y.-W.C.; project administration, Y.-W.C. and J.S.; funding acquisition, J.S. and Y.-W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Natural Science Foundation of Xiamen City, Fujian Province, China, under the Grant No. 3502ZZ20227199 and in part by the Grant-in-Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture, and Sports (MEXT) under the Grant No. 20KK0234 and No. 21H03470.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Ritsumeikan University under No. BKC-LSMH-2021-037.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments: The authors would like to thank Hou Ruibo of Ritsumeikan University, Japan, for his helpful advice on this research and Rahul Jain of Ritsumeikan University, Japan, for his kind English proof.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Radiology Assistant, “CT Contrast Injection and Protocols,” Radiology Dept. Rijnland Hospital, Leiderdorp, the Netherlands. [Online]. Available online: <http://www.radiologyassistant.nl/en/p52c04470dbd5c/ct-contrast-injection-and-protocols.html> (accessed on 19 October 2025).
2. Yasaka, K.; Akai, H.; Abe, O.; Kiryu, S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. *Radiology* **2018**, *286*, 887–896.
3. Liang, D.; Liu, M.; Zhang, J.; Wang, Y.; Zhang, D. Combining convolutional and recurrent neural networks for classification of focal liver lesions in multi-phase CT imaging. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, 2nd ed.; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Cham, Switzerland, 2018; Volume 11071, pp. 666–675.
4. Wang, W.; Wang, Y.; Liang, D.; Zhang, J. Classification of focal liver lesions using deep learning with fine-tuning. In *Proceedings of the Digital Medicine and Image Processing (DMIP2018)*, Chengdu, China, 24–26 November 2018; pp. 56–60.
5. Dong, H.; Iwamoto, Y.; Han, X.H.; Lin, L.; Hu, H.; Cai, X.; Chen, Y.-W. Case Discrimination: Self-supervised Feature Learning for the Classification of Focal Liver Lesions. In *Innovation in Medicine and Healthcare, Proceedings of the 9th KES-InMed 2021, Virtual Event, 14–16 June 2021*; Chen, Y.-W., Tanaka, S., Howlett, R.J., Jain, L.C., Eds.; Springer: Singapore, 2021; Volume 254, pp. 241–249.
6. Song, J.; Dong, H.; Chen, Y.; Lin, L.; Hu, H.; Chen, Y.-W. Deep Neural Network-Based Classification of Focal Liver Lesions Using Phase-Shuffle Prediction Pre-training. In *Innovation in Medicine and Healthcare, Proceedings of the 11th KES-InMed 2023, Rome, Italy, 14–16 June 2023*; Chen, Y.-W., Tanaka, S., Howlett, R.J., Jain, L.C., Eds.; Smart Innovation, Systems and Technologies; Springer: Cham, Switzerland, 2023; Volume 357, pp. 235–243.
7. Song, J.; Dong, H.; Chen, Y.; Zhang, X.; Zhan, G.; Jain, R.K.; Chen, Y.-W. Early Recurrence Prediction of Hepatocellular Carcinoma Using Deep Learning Frameworks with Multi-Task Pre-Training. *Information* **2024**, *15*, 493.
8. Desai, K.; Johnson, J. VirTex: Learning Visual Representations from Textual Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Nashville, TN, USA, 19–25 June 2021; IEEE: New York, NY, USA, 2021; pp. 11162–11172.

9. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv* **2020**, arXiv:2010.00747.
10. Hayat, M.; Aramvith, S.; Bhattacharjee, S.; Ahmad, N. Attention GhostUNet++: Enhanced Segmentation of Adipose Tissue and Liver in CT Images. *arXiv* **2025**, arXiv:2504.11491.
11. Bawazir, A.; Wu, K.; Li, W. Uni-Mlip: Unified Self-supervision for Medical Vision Language Pre-training. *arXiv* **2024**, arXiv:2411.15207.
12. Song, J.; Hu, Y.; Wang, H.; Chen, Y.-W. Liver-VLM: A Vision-Language Model for Focal Liver Lesion Classification. In Proceedings of the 2025 International Conference on Innovation in Medicine and Healthcare, Solin, Croatia, 25–27 June 2025. <https://doi.org/10.48550/arXiv.2505.03350>.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 770–778.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual Event, 3–7 May 2021.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
16. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
17. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP 2019), Minneapolis, MN, USA, 7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 72–78.
18. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270.
19. Available online: <https://github.com/google-research/bert> (accessed on 19 October 2025).
20. Xu, Y.; Zhou, H.; Zhang, Z.; Wang, Y.; Xie, Y. PA-ResSeg: A Phase Attention Residual Network for Liver Tumor Segmentation from Multi-phase CT Images. *Med. Phys.* **2021**, *48*, 3752–3766.
21. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
22. Chen, X.; Li, X.; Fan, H. Exploring Simple Siamese Representation Learning. *arXiv* **2020**, arXiv:2011.10566.
23. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. *arXiv* **2021**, arXiv:2106.13230.
24. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. *arXiv* **2022**, arXiv:2201.03545.
25. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. *arXiv* **2021**, arXiv:2104.00298.
26. Student. The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1–25.
27. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837–845.
28. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv* **2016**, arXiv:1610.02391.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.