FedCHDP: Federated Concept-Guided Dual-Modal Prompting for CHD Diagnosis

Ruilin Gu[1,*], He Li[1,*], Wenke Huang[2], Qingxiong Tan[1,3],

Mang Ye[1,4,†], and Bo Du[1,†]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Nanyang Technological University, Singapore
[3] Institute for Math and AI, Wuhan University, Wuhan, China
[4] Taikang Center for Life and Medical Sciences, Wuhan University, Wuhan, China

Abstract. Early detection of Congenital Heart Disease (CHD) relies heavily on subjective ultrasound interpretation. While AI offers automation, privacy regulations and data fragmentation remain major barriers. Federated Learning (FL) enables privacy-preserving collaboration, yet deploying pre-trained vision-language models like CLIP in non-IID settings is challenged by inter-hospital heterogeneity and semantic overfitting to site-specific biases. To address these issues, we propose FedCHDP, a novel federated concept-guided dual-modal prompting framework. On the textual side, we construct a federated medical concept library that decomposes diagnostic queries into fine-grained clinical descriptors (i.e., chamber numbers, flow numbers and flow patterns), enforcing semantic consensus across clients. An uncertainty-aware gating mechanism further adaptively fuses dynamic concept cues with global priors to rectify ambiguous predictions. On the visual side, we introduce a global-guided contrastive alignment that calibrates visual prompts using disease prototypes, aligning local embeddings with global semantic centers to mitigate distribution shifts. Experiments show that FedCHDP significantly outperforms state-of-the-art methods for multi-center CHD identification.

Keywords: Federated Learning · Congenital Heart Disease.

1 Introduction

Congenital Heart Disease (CHD) is the most prevalent birth defect worldwide [28, 15]. While early detection is critical for neonatal outcomes, current prenatal screening primarily relies on ultrasound assessment during the **second trimester** (18–22 weeks) [24, 3, 9], missing the optimal intervention window and raising ethical concerns if pregnancy termination is necessary [5]. Advancing screening to the **first trimester** (11–14 weeks) is therefore a clinical imperative. However, the fetal heart's minute, complex structure, coupled with subjective expert interpretation, poses significant diagnostic challenges [1, 2]. While Artificial

† Corresponding Author {yemang, dubo}@whu.edu.cn

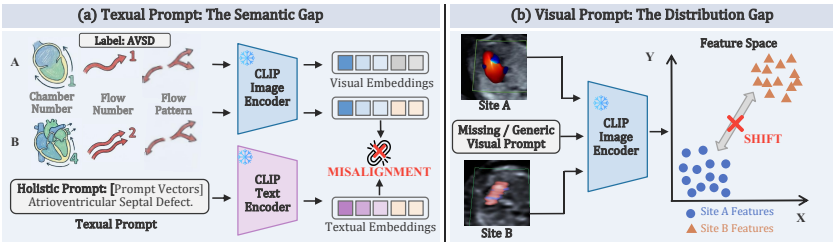

Fig. 1: Challenges in federated medical prompt tuning. (a) **Semantic Gap:** holistic prompts fail to capture varying anatomical concepts (e.g., chamber/flow numbers) within the same disease category, limiting diagnostic precision. (b) **Distribution Gap:** Severe heterogeneity across institutions causes feature divergence, creating a fragmented semantic space that hampers global generalization.

Intelligence (AI) offers promise for standardizing early screening [17], developing robust models requires large-scale data that is often inaccessible due to privacy regulations [18, 6], creating a significant barrier rooted in **data fragmentation**.

Federated Learning (FL) addresses this by enabling multi-center collaborative training without sharing raw patient data [21, 4, 11, 22, 25]. Pre-trained Vision-Language models such as CLIP [20], combined with Parameter-Efficient Fine-Tuning (PEFT) techniques [7], have become pivotal in medical imaging for their superior generalization and low computational cost. However, as illustrated in Figure 1, deploying these generic frameworks in federated medical settings introduces two critical challenges. **I) The Semantic Ambiguity and Individual Variability**: Existing textual prompt methods rely on holistic, opaque learnable vectors that fail to capture universal medical semantics from fragmented local data [27, 12]. This is especially problematic in CHD diagnosis, where even patients sharing the same diagnosis, such as Atrioventricular Septal Defect, can exhibit vastly different anatomical configurations[16, 13]. Compressing such diverse pathology into a single coarse-grained vector prevents instance-level adaptation, ultimately limiting global model performance. **II) The Distribution Shift and Visual Bias:** Federated environments exhibit inherent heterogeneity in imaging protocols, equipment, and operator skill [21, 8]. This induces severe feature shifts that cause visual prompts to overfit local imaging characteristics rather than learning universal pathological features, producing inconsistent cross-site representations and degrading model robustness [23, 14].

To address both challenges, we propose **FedCHDP**, a federated medical concept-guided dual-modal prompting framework. To resolve semantic ambiguity and intra-class variability, we introduce a **federated medical concept-driven prompting** strategy that replaces unstructured prompts with fine-grained clinical descriptors (e.g., chamber numbers, blood flow counts, and flow patterns) drawn from a shared medical concept library. This generates instance-adaptive textual representations that preserve individual anatomical detail while maintaining global semantic consistency. Inspired by clinical reasoning, we further incorporate an **uncertainty-aware gating mechanism** that dynamically bal-

ances global disease priors with instance-specific anatomical cues, enabling reliable prediction refinement in complex cases. To mitigate cross-center distribution shifts, we introduce a **global-guided contrastive alignment** that regularizes visual prompt learning by aggregating decentralized features into global disease prototypes. Local embeddings are aligned with their corresponding prototypes while being repelled from mismatched subtypes, promoting domain-consistent representations and suppressing site-specific overfitting.

Our main contributions are summarized as follows:

1. We establish a new privacy-preserving paradigm for **first-trimester collaborative CHD screening**, demonstrating how shared semantic consensus enables large vision-language models to be efficiently adapted for early-stage prenatal diagnosis under strict privacy constraints.
2. We present **FedCHDP**, a unified framework that leverages **dual-modal prompting**. It adaptively models individual anatomical variations via a textual concept library, and calibrates visual prompt tuning through global prototype contrastive alignment to eliminate site-specific biases.
3. Extensive evaluations on a multi-center real-world dataset demonstrate that FedCHDP significantly **outperforms state-of-the-art methods**, validating its robustness and potential for clinical deployment.

2 Methodology

2.1 Preliminaries and Problem Formulation

Given K decentralized clients with private datasets $\mathcal{D}_k = \{(\mathbf{x}_i, y_i, \mathbf{c}_i)\}_{i=1}^{N_k}$, where $\mathbf{x}_i$ represents a first-trimester fetal cardiac ultrasound image, $y_i \in \{1, \ldots, C\}$ denotes the CHD subtype, and $\mathbf{c}_i = \{c_{i,m}\}_{m \in \{ch, fl, pa\}}$ represents fine-grained clinical concepts, including chamber number (ch), flow number (fl), and flow pattern (pa), respectively. For each dimension m, K_m denotes the number of possible discrete concept values, and $c_{i,m} \in \{1, \ldots, K_m\}$ is the ground-truth discrete concept label of sample i for concept m. For example, the flow number concept has two possible values, i.e., one flow or two flows, and thus $K_{fl} = 2$.

We build our framework upon frozen pre-trained CLIP visual ($\mathcal{V}$) and textual ($\mathcal{T}$) encoders, and adapt them via parameter-efficient prompt tuning. Specifically, a learnable visual prompt $\mathbf{P}_v$ is injected into the frozen visual encoder together with the input image $\mathbf{x}_i$, producing the visual embedding $\mathbf{v}_i = \mathcal{V}(\mathbf{x}_i, \mathbf{P}_v)$. Meanwhile, unconditional and concept-conditioned textual prompts are encoded by $\mathcal{T}$ and matched with $\mathbf{v}_i$, producing logits $\mathbf{l}_{up}$ and $\mathbf{l}_{ccp}$. These logits are adaptively fused into a final prediction $\mathbf{l}_{fused}$ via an uncertainty-aware gate, optimized by the classification loss $\mathcal{L}_{cls} = \mathrm{CE}(\mathbf{l}_{fused}, y_i)$.

2.2 Textual Side: Concept-Guided Dynamic Prompting

Dual-Branch Prompt Construction We propose a dual-branch structure: an *unconditional disease prompt* ($\mathbf{P}_{up}$) for global consensus and a *concept-conditioned prompt* ($\mathbf{P}_{ccp}$) for instance-adaptive specialization.

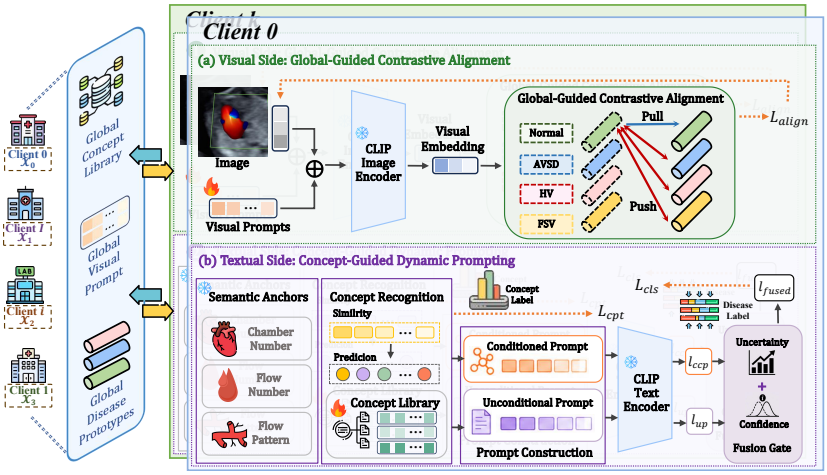

Fig. 2: Overview of the proposed dual-modal prompting framework, FedCHDP: (a) a globally guided contrastive alignment **visual branch**, which calibrates visual prompts using shared disease prototypes; and (b) a concept-guided dynamic prompting **textual branch**, which constructs and fuses unconditional and concept-conditioned prompts via an uncertainty-aware gating mechanism.

We decompose CHD-related anatomical knowledge into three clinically interpretable concept dimensions: chamber number, flow number, and flow pattern. For each dimension m, we maintain a *federated medical concept library* $\mathcal{E}_m = \{\mathbf{e}_{m,j}\}_{j=1}^{K_m}$, where $\mathbf{e}_{m,j}$ is a learnable prompt embedding corresponding to the j-th concept value of dimension m and is collaboratively optimized across clients. To stabilize concept recognition against cross-center domain heterogeneity, we introduce non-trainable *semantic consensus anchors* $\mathcal{A}_m = \{\mathbf{a}_{m,j}\}_{j=1}^{K_m}$. These anchors are generated by feeding standardized clinical text templates (e.g., two blood flows) into the frozen text encoder $\mathcal{T}$, thereby providing stable and semantically meaningful references for concept recognition. Given the visual embedding $\mathbf{v}_i$, we compute the scaled cosine similarities $\mathbf{s}_{i,m}$ between $\mathbf{v}_i$ and each semantic anchor in concept m. The predicted concept $\hat{c}_{i,m} = \arg\max_{j\in\{1,\dots,K_m\}} \mathbf{s}_{i,m}^{(j)}$ is supervised by ground-truth discrete concept label $c_{i,m}$ via $\mathcal{L}_{cpt} = \sum_m \mathrm{CE}(\mathbf{s}_{i,m}, c_{i,m})$.

For each individual i, the predicted concept $\hat{c}_{i,m}$ is used to retrieve the corresponding learnable concept embedding $\mathbf{e}_{m,\hat{c}_{i,m}}$ from $\mathcal{E}_m$. Based on the retrieved embeddings from all concept dimensions, we construct a concept-conditioned prompt for each disease class c: $\mathbf{P}_{ccp}^c = [\mathbf{e}_{dis}; \mathbf{e}_{ch,\hat{c}_{i,ch}}; \mathbf{e}_{fl,\hat{c}_{i,fl}}; \mathbf{e}_{pa,\hat{c}_{i,pa}}; \mathbf{w}_c]$, where $\mathbf{e}_{dis}$ and $\mathbf{w}_c$ denote the shared disease context and class token. The prompt $\mathbf{P}_{ccp}^c$ is encoded by the frozen text encoder $\mathcal{T}$, and its image-text similarity with $\mathbf{v}_i$ produces the concept-conditioned logit $\mathbf{l}_{ccp}$. Concurrently, we construct an unconditional disease prompt by replacing the instance-specific concept embeddings with concept-agnostic generic tokens: $\mathbf{P}_{up}^c = [\mathbf{e}_{dis}; \mathbf{e}_{ch}^u; \mathbf{e}_{fl}^u; \mathbf{e}_{pa}^u; \mathbf{w}_c]$, producing stable logits $\mathbf{l}_{up}$, where $\mathbf{e}_m^u$ is a learnable generic token for concept dimension m.

Uncertainty-Aware Fusion To emulate expert diagnostic reasoning, we activate concept-conditioned refinement only when the unconditional disease prediction is uncertain and the inferred anatomical evidence is reliable. For the i-th individual, we quantify disease-level uncertainty by normalized predictive entropy $u_i = H(\mathbf{p}_i)/\log C$, where $\mathbf{p}_i = \text{Softmax}(\mathbf{l}_{up})$. For each concept dimension m, we compute the concept probability distribution $\mathbf{q}_{i,m} = \text{Softmax}(\mathbf{s}_{i,m})$. The concept confidence is defined as $conf_{i,m} = \frac{1}{2}(q_{i,m}^{(1)} - q_{i,m}^{(2)}) + \frac{1}{2}(1 - H(\mathbf{q}_{i,m})/\log K_m)$, where $q_{i,m}^{(k)}$ denotes the k-th highest probability in $\mathbf{q}_{i,m}$. The overall anatomical reliability is aggregated via the geometric mean $conf_i = (conf_{i,ch} \cdot conf_{i,fl} \cdot conf_{i,pa})^{1/3}$. The final prediction uses a gating coefficient $g_i = \min(u_i \cdot conf_i, 1)$:

$$\mathbf{l}_{fused} = \mathbf{l}_{up} + g_i \cdot (\mathbf{l}_{ccp} - \mathbf{l}_{up}). \tag{1}$$

This mechanism rectifies challenging cases using patient-specific anatomical details while maintaining stability for routine samples.

2.3 Visual Side: Global-Guided Contrastive Alignment

Hierarchical Prototype Construction To encourage domain-consistent visual representations, we introduce a global-guided contrastive alignment based on disease prototypes. At the end of each local training round, client k computes a local prototype $\mu_{k,c}$ for every disease subtype $c \in \mathcal{Y}$ by averaging local visual embeddings: $\mu_{k,c} = \frac{1}{|\mathcal{D}_{k,c}|} \sum_{i \in \mathcal{D}_{k,c}} \mathbf{v}_i$, where $|\mathcal{D}_{k,c}|$ denotes the set of samples from client k with label c. The server then aggregates prototypes class-wise with sample-size weighting to mitigate noise from scarce local class samples:

$$\mu_c^{global} = \frac{\sum_{k=1}^{K} |\mathcal{D}_{k,c}| \cdot \mu_{k,c}}{\sum_{k=1}^{K} |\mathcal{D}_{k,c}|}. \tag{2}$$

Contrastive Visual Calibration After aggregation, the global prototypes are broadcast back to clients and serve as disease-level visual references. During local training, we utilize a prototype-based contrastive objective to pull the visual embedding $\mathbf{v}_i$ toward its corresponding positive global prototype $\mu_{y_i}^{global}$ and push it away from mismatched prototypes $\mu_{j \neq y_i}^{global}$:

$$\mathcal{L}_{align} = -\log \frac{\exp(\langle \mathbf{v}_i, \mu_{y_i}^{global} \rangle / \tau_v)}{\sum_{j=1}^{C} \exp(\langle \mathbf{v}_i, \mu_j^{global} \rangle / \tau_v + \epsilon)}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, τ_v is the temperature, and ϵ ensures numerical stability. The loss calibrates the visual representation space by constraining the learnable visual prompt, effectively mitigating cross-center distribution shifts and promoting a separable, domain-consistent representation space.

Overall Loss Function The overall training objective integrates three components to ensure robust and effective multi-modal learning: the classification loss $\mathcal{L}_{cls}$ for disease prediction, the concept supervision loss $\mathcal{L}_{cpt}$ for accurate anatomical concept inference, and the contrastive alignment loss $\mathcal{L}_{align}$ for visual representation learning. The final optimization objective is formulated as

Table 1: **Quantitative comparison** of federated prompt learning methods on multi-class CHD dataset. **Bold** and underline denote the best and second-best performance, respectively. The ↑ and ↓ indicate the absolute improvement or degradation compared with the best baseline method. A dash (–) is used where specificity is considered uninformative due to zero sensitivity.

Method	Overall			Normal			AVSD			FSV			HV		
	F1	SENS	SPEC	F1	SENS	SPEC	F1	SENS	SPEC	F1	SENS	SPEC	F1	SENS	SPEC
ZS-CLIP[20]	7.19	22.40	74.59	0.00	0.00	–	1.67	1.00	97.00	20.74	**84.24**	7.59	6.37	4.37	93.78
CoCoOp[27]	42.68	45.94	86.23	86.47	93.42	69.71	41.90	32.84	96.11	37.34	54.89	79.54	5.00	2.62	**99.59**
KgCoCoOp[26]	29.58	36.67	83.89	83.66	90.13	66.12	0.00	0.00	–	30.63	54.35	70.50	4.05	2.18	98.92
FedTPG[19]	39.47	42.41	85.31	85.07	91.23	68.89	7.27	3.98	**99.11**	41.83	52.17	85.69	23.72	22.27	87.56
PromptFL[7]	44.92	47.97	86.69	87.04	94.52	69.71	37.74	29.85	95.38	43.88	61.41	82.57	11.02	6.11	99.08
FedVPT[10]	52.28	55.10	89.10	90.57	96.47	77.85	41.38	35.82	93.92	55.20	75.00	85.77	21.98	13.10	98.84
LPT[12]	44.64	47.68	86.29	85.97	93.30	68.24	39.88	32.84	94.81	44.27	59.78	83.77	8.46	4.80	98.34
IVLP[12]	57.91	59.53	90.96	93.18	95.62	**87.13**	43.99	31.84	97.89	51.83	80.98	80.66	42.63	29.69	98.18
MaPLe[12]	55.97	58.62	90.36	92.30	96.35	83.39	46.29	38.81	95.30	56.12	80.98	84.17	29.17	18.34	98.59
FedCHDP	**67.22**	**66.73**	**92.22**	**93.67**	**98.29**	84.53	**53.74**	**39.30**	98.87	**60.68**	82.61	**86.81**	**60.80**	**46.72**	98.67
	↑9.31	↑7.20	↑1.26	↑0.49	↑1.82	↓2.6	↑7.45	↑0.49	↓0.24	↑4.56	↓1.63	↑1.04	↑18.17	↑17.03	↓0.92

$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{cpt} \cdot \mathcal{L}_{cpt} + \lambda_{align} \cdot \mathcal{L}_{align}$, where λ_{cpt} and λ_{align} balance the contributions of concept guidance and visual contrastive alignment, respectively.

3 Experiments

3.1 Experimental Setup

We evaluate the proposed FedCHDP framework on a multi-center dataset consisting of 5,108 first-trimester fetal echocardiography images sourced from four tertiary hospitals. Following strict ethical approval, domain experts annotated these images for structural regions and explicit clinical concepts, specifically chamber numbers, flow numbers, and flow patterns. The dataset is categorized into four distinct classes: Normal, Atrioventricular Septal Defect (AVSD), Hypoplastic Ventricle (HV), and Functional Single Ventricle (FSV). To robustly evaluate diagnostic performance across this heterogeneous data, we employ Macro-F1, Sensitivity, and Specificity as our primary metrics.

Ethical approval was obtained from the local ethics committee at each participating center. Approval was granted by the ethics committee of BOGH(No. 2022-KY-097-03), GZAH(No. 20194), XH(No. 202211714), CH(No. CSSFYBJYEC-SOP-005-F07V1.0). Written informed consent was obtained from participants prior to data collection. The study was approved by the institutional review boards of all participating centers (ChiCTR2400086975) in accordance with the Declaration of Helsinki and follows the TRIPOD reporting guidelines.

We utilize a frozen CLIP (ViT-B/16) backbone. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU using FP32 precision to ensure numerical stability. Input images are resized to 224×224 pixels and augmented via random cropping and horizontal flipping. We set the prompt length

Table 2: **Ablation analysis** of the proposed components on both textual and visual sides, validating the effectiveness of concept-guided prompting and global contrastive alignment.

Textual	Visual	Overall	Normal	AVSD	FSV	HV
✗	✗	57.04	95.30	52.26	50.59	30.03
✗	✓	62.80	91.53	45.45	59.08	55.14
✓	✗	60.17	**96.33**	48.23	54.39	41.72
✓	✓	**67.22**	93.67	**53.74**	**60.68**	**60.80**

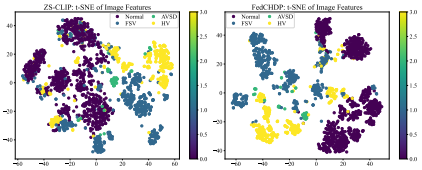

Fig. 3: **t-SNE comparison** of ZS-CLIP vs. FedCHDP. Our method achieves superior category-wise clustering and distinct decision boundaries..

to 16, and the shared concept embedding dimension to 512. To maintain parameter efficiency, learnable visual prompts are restricted to the first encoder layer. The model undergoes federated training for 50 global communication rounds, with each client executing 5 local epochs per round. Optimization is performed using the SGD optimizer with a learning rate of 1e-3 and a batch size of 32. Additionally, the temperature parameter for global-guided contrastive alignment is fixed at 0.5, and the dataset is partitioned using a standard 7:3 train-test split.

To demonstrate the superiority of our approach, we conduct comprehensive benchmark evaluations of FedCHDP against a strong and diverse set of federated prompt learning methods. Moreover, we incorporate two representative comparison strategies originally proposed in MaPLe [12]. Specifically, LPT introduces learnable prompts into Transformer layers of the language encoder to progressively adapt textual representations, while IVLP jointly learns prompts in both vision and language branches without cross-modal interaction, enabling a more comprehensive and fair comparison.

3.2 Results

Comparison with State-of-the-Art Methods Table 1 presents the comprehensive quantitative results. Our proposed FedCHDP significantly outperforms all baselines, achieving state-of-the-art Overall F1, Sensitivity, and Specificity by massive margins (e.g., a +9.31% absolute F1 improvement over the strongest baseline, IVLP). Notably, Zero-Shot CLIP struggles with the vast natural-to-medical domain gap. While other federated prompt learning methods improve upon this, their robustness remains limited, relying on holistic semantic representations, they fail to capture the fine-grained anatomical nuances crucial for CHD subtyping. Some models completely fail to recognize certain disease subtypes (e.g., AVSD), yielding zero sensitivity and F1-score, this reveals severe performance degradation and an inability to detect positive cases for these categories. In the critical context of early prenatal CHD screening, maintaining high Sensitivity is paramount, as missing a structural defect carries severe medical consequences. FedCHDP eliminates this trade-off, demonstrating highly balanced and robust diagnostic capabilities across all challenging rare subclasses (AVSD, FSV, and HV) without sacrificing overall reliability.

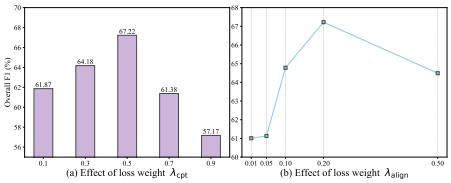

Fig. 4: **Hyper-parameter analysis.** We analyze the loss weight λ_{cpt} and λ_{align}. The results indicate that the model achieves satisfying performance when $\lambda_{cpt} = 0.5$ and $\lambda_{align} = 0.2$.

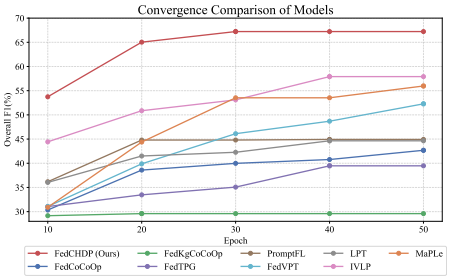

Fig. 5: **Convergence comparison.** FedCHDP demonstrates superior convergence speed and the highest peak F1 score across 50 epochs.

Ablation Study To validate the individual contributions of our core modules, we conducted an ablation study (Table 2). The baseline employing standard dual-modal prompt tuning achieves a modest Overall F1 of 57.04%. Incorporating only the Concept-Guided Textual Prompting improves the Overall F1 to 60.17%, demonstrating that explicitly grounding prompts in clinical dimensions effectively resolves semantic ambiguity. Applying only the Global-Guided Contrastive Alignment on the visual branch yields a substantial boost to 62.80%. Notably, this visual alignment drastically improves the recognition of challenging minority classes, such as HV (from 30.03% to 55.14%) and FSV (from 50.59% to 59.08%), confirming its ability to mitigate site-specific domain shifts and prevent local overfitting. Finally, our complete FedCHDP synergizes both modalities, achieving the highest Overall F1 of 67.22% and optimal performance across the most difficult CHD subtypes.

Qualitative Analysis To intuitively illustrate the representational capability of our framework, we visualize the image features using t-SNE in Fig. 3. While ZS-CLIP features are highly entangled with overlapping classes, FedCHDP produces a discriminative embedding space characterized by strong intra-class compactness and clear inter-class separability. This demonstrates that our prompting strategy effectively captures domain-invariant, disease-specific representations despite the heterogeneity of the federated environment.

Hyperparameter Sensitivity and Convergence To identify the optimal hyperparameter configuration, we employ a control variable strategy, evaluating model performance by systematically varying one parameter while keeping the others constant. As shown in Fig. 4, the overall F1-score exhibits an initial upward trend followed by a decline as λ_{cpt} and λ_{align} increase. Under this empirical trend, the model achieves its peak performance at $\lambda_{cpt} = 0.5$ and $\lambda_{align} = 0.2$. Furthermore, the convergence analysis in Fig. 5 shows that FedCHDP converges faster and reaches a higher F1-score than competing methods, demonstrating superior optimization efficiency under federated heterogeneity.

4 Conclusion

We introduced FedCHDP, a novel federated dual-modal prompting framework for privacy-preserving, first-trimester CHD screening. To bridge the semantic gap caused by inter-patient variability of complex CHD phenotypes, the textual branch establishes a shared medical concept library that explicitly grounds textual prompts in fine-grained clinical descriptors. Furthermore, to overcome the distribution gap arising from severe cross-center data heterogeneity, the visual branch employs a global-guided contrastive alignment module to calibrate visual prompt tuning using aggregated disease prototypes. Extensive evaluations on a real-world, multi-center dataset demonstrate that FedCHDP significantly outperforms state-of-the-art methods. Crucially, it achieves robust classification on highly challenging, rare CHD subtypes without model collapse, alongside faster convergence. By bridging both gaps, FedCHDP provides a reliable, generalizable, and clinically viable solution for decentralized prenatal diagnosis.

Acknowledgments. This work is partially funded by the National Key Research Development Program of China (2023YFC2705702, 2026YFE0202100), the National Natural Science Foundation of China under Grant (62361166629, 62506269, 623B2080), the Innovative Research Group Project of Hubei Province under Grants (2024AFA017), and the General Program of Hubei Provincial Natural Science Foundation under Grant (JCZRMS202600468). This work was also supported by the WHU-Kingsoft Joint Lab. The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper.

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

References

1. Arnaout, R., Curran, L., Zhao, Y., Levine, J.C., Chinn, E., Moon-Grady, A.J.: An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nature Medicine **27**(5), 882–891 (2021)
2. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
3. Carvalho, J.S., Allan, L., Chaoui, R., Copel, J., DeVore, G., Hecher, K., Lee, W., Munoz, H., Paladini, D., Tutschek, B., Yagel, S.: ISUOG Practice Guidelines (updated): sonographic screening examination of the fetal heart. Ultrasound in Obstetrics & Gynecology **41**(3), 348–359 (2013)
4. Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al.: Federated learning for predicting clinical outcomes in patients with covid-19. Nature Medicine **27**(10), 1735–1743 (2021)
5. Eckersley, L., Sadler, L., Parry, E., Finucane, K., Gentles, T.L.: Timing of diagnosis affects mortality in critical congenital heart disease. Archives of Disease in Childhood **101**(6), 516–520 (2016)

6. European Parliament and Council of the European Union: General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016)
7. Guo, T., Guo, S., Wang, J., Tang, X., Xu, W.: Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. IEEE Transactions on Mobile Computing **23**(5), 5179–5194 (2023)
8. Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16312–16322. IEEE (2023)
9. Hunter, L.E., Simpson, J.M.: Prenatal screening for structural congenital heart disease. Nature Reviews Cardiology **11**(6), 323–334 (2014)
10. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
11. Kang, M., Chikontwe, P., Kim, S., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H.: One-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 521–531. Springer (2023)
12. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19113–19122 (2023)
13. Kong, F., Stocker, S., Choi, P.S., Ma, M., Ennis, D.B., Marsden, A.L.: SDF4CHD: Generative modeling of cardiac anatomies with congenital heart defects. Medical Image Analysis **97**, 103293 (2024)
14. Ma, Y., Dai, W., Huang, W., Chen, J.: Geometric knowledge-guided localized global distribution alignment for federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20958–20968 (2025)
15. Marelli, A.J., Ionescu-Ittu, R., Mackie, A.S., Guo, L., Dendukuri, N., Kaouache, M.: Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. Circulation **130**(9), 749–756 (2014)
16. Micheletti, A.: Congenital heart disease classification, epidemiology, diagnosis, treatment, and outcome. In: Congenital Heart Disease: the Nursing Care Handbook, pp. 1–67. Springer (2018)
17. Nurmaini, S., Partan, R.U., Bernolian, N., Sapitri, A.I., Tutuko, B., Rachmatullah, M.N., Darmawahyuni, A., Firdaus, F., Mose, J.C.: Deep learning for improving the effectiveness of routine prenatal screening for major congenital heart diseases. Journal of Clinical Medicine **11**(21), 6454 (2022)
18. Price, W.N., Cohen, I.G.: Privacy in the age of medical big data. Nature Medicine **25**(1), 37–43 (2019)
19. Qiu, C., Li, X., Mummadi, C.K., Ganesh, M.R., Li, Z., Peng, L., Lin, W.Y.: Federated text-driven prompt generation for vision-language models. In: The Twelfth International Conference on Learning Representations. pp. 12997–13012 (2024)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
21. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ Digital Medicine **3**(1), 119 (2020)

22. Sahoo, P., Tripathi, A., Saha, S., Mondal, S.: Fedmrl: Data heterogeneity aware federated multi-agent deep reinforcement learning for medical imaging. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 640–649. Springer (2024)
23. Shao, J., Wu, F., Zhang, J.: Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. Nature Communications **15**(1), 349 (2024)
24. Tegnander, E., Eik-Nes, S.: The examiner's ultrasound experience has a significant impact on the detection rate of congenital heart defects at the second-trimester fetal examination. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology **28**(1), 8–14 (2006)
25. Yan, Y., Zhu, L., Li, Y., Xu, X., Goh, R.S.M., Liu, Y., Khan, S., Feng, C.M.: A new perspective to boost performance fairness for medical federated learning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 13–23. Springer (2024)
26. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6757–6767 (2023)
27. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16816–16825 (2022)
28. Zimmerman, M.S., Smith, A.G.C., Sable, C.A., Echko, M.M., Wilner, L.B., Olsen, H.E., Atalay, H.T., Awasthi, A., Bhutta, Z.A., Boucher, J.L., et al.: Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the global burden of disease study 2017. The Lancet Child & Adolescent Health **4**(3), 185–200 (2020)