TPA: Temporal Prompt Alignment for Fetal Congenital Heart Defect Classification

Darya Taratynova, Alya Almsouti, Beknur Kalmagambet, Numan Saeed, & Mohammad Yaqub
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

darya.taratynova@mbzuai.ac.ae

Paper

Code

INTRODUCTION

- **CHD** affects **~1% of live births**; many cases need surgery.
- **Early, accurate detection** is vital for better outcomes.
- Ultrasound is standard but is limited by **operator dependence, image artifacts**

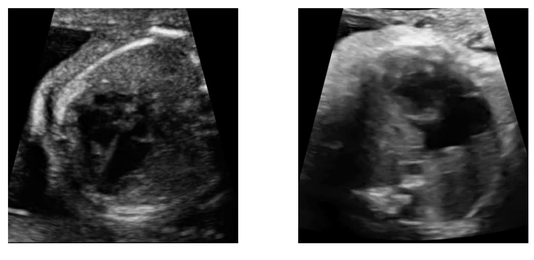


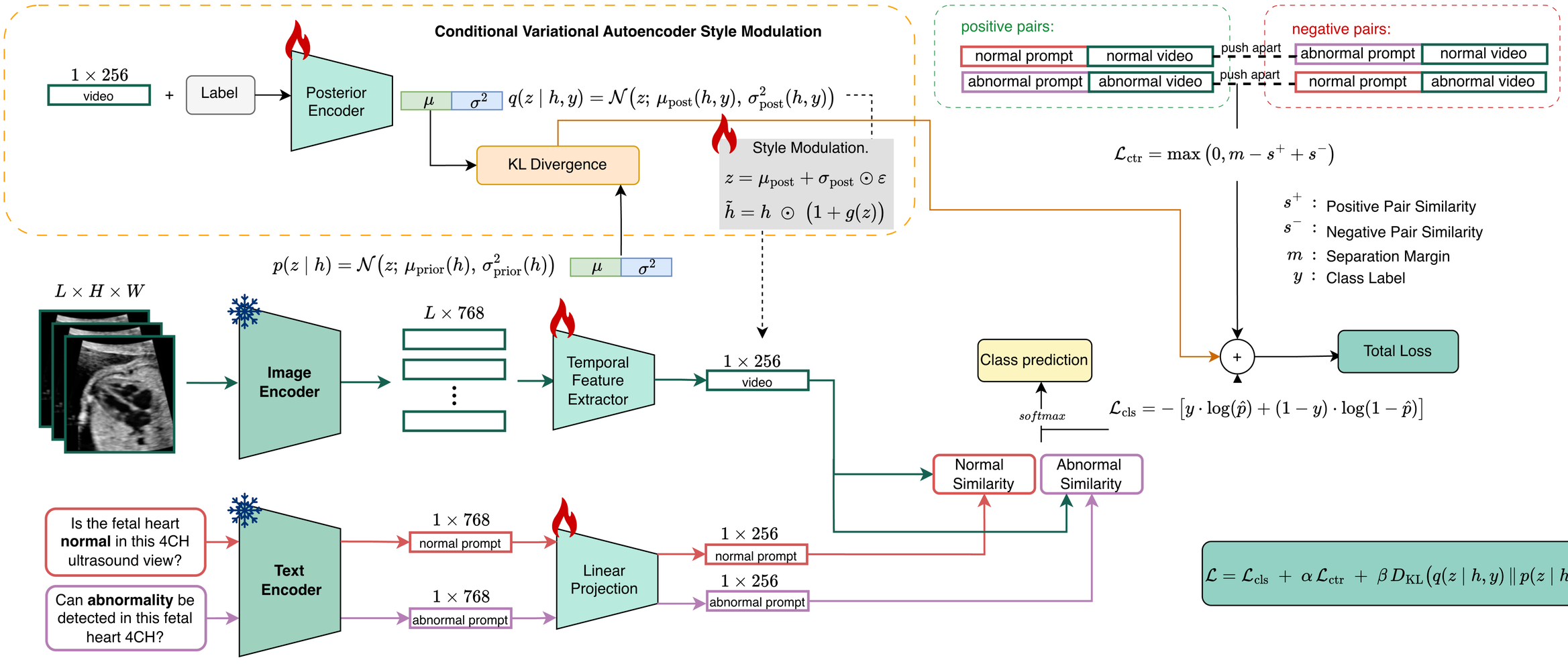
Fig 1. Fetal 4CH ultrasound showing Speckle Noise (left) and Motion (right)

CONTRIBUTIONS

- We propose a novel classification framework that combines vision–language foundation models with trainable temporal modeling. By integrating a lightweight temporal extractor and a **prompt-aware contrastive loss**, our method effectively captures spatiotemporal dynamics and aligns video content with text prompts.
- We propose the first **Conditional Variational Autoencoder with Style Modulation for video classification**, operating on temporal video embeddings to enable reliable video-level uncertainty estimation.

METHODOLOGY

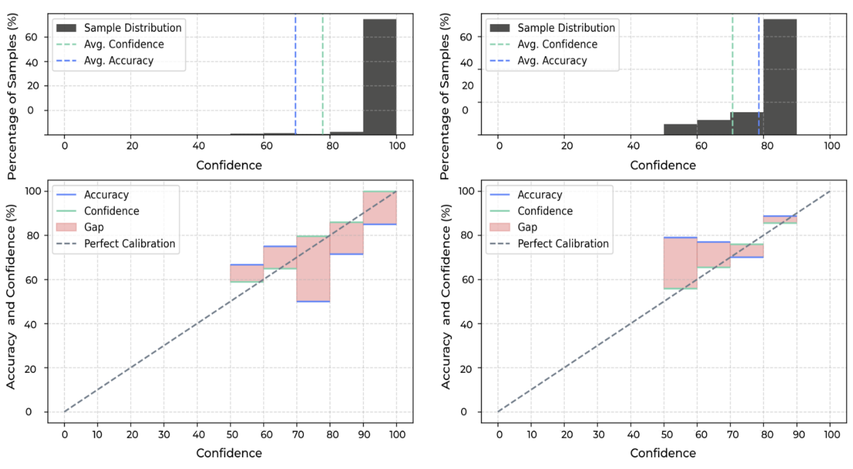
Fig 2. TPA with Conditional Variational Autoencoder Style Modulation



RESULTS

Model	F1 (↑)	AUC (↑)
Frame-wise	73.98 ± 2.89	76.16 ± 2.87
TCN [23]	81.83 ± 1.94	85.44 ± 3.98
xLSTM [4]	80.32 ± 1.99	76.16 ± 2.87
GNN [7]	84.84 ± 1.64	87.69 ± 3.06
⋮		
TPA (ours)	**85.40 ± 2.11**	**88.31 ± 4.43**

Table 1. Binary Classification of CHD



Method	F1 (↑)	AECE (↓)	ECE (↓)
Before	**85.40 ± 2.36**	39.76 ± 4.81	15.98 ± 3.83
After CVAESM	84.12 ± 1.52	**32.91 ± 4.49**	**10.60 ± 4.40**

Fig 3. Calibration before and after applying CVAESM

Class	TCN	GNN	xLSTM	TPA
Defects = 2				
Macro	74.23 ± 9.58	72.33 ± 6.98	75.42 ± 7.50	**78.68 ± 8.20**
Normal	85.62 ± 6.99	86.23 ± 6.70	86.71 ± 4.99	**88.69 ± 5.24**
VSD	44.12 ± 1.80	52.27 ± 19.89	43.88 ± 17.57	**53.96 ± 1.13**
AVSD	61.04 ± 19.78	57.19 ± 18.32	**61.06 ± 19.35**	59.56 ± 2.06
Defects = 3				
Macro	68.14 ± 5.40	70.65 ± 5.64	64.79 ± 4.51	**72.91 ± 3.75**
Normal	82.83 ± 4.32	85.10 ± 3.00	**85.54 ± 3.62**	84.62 ± 2.47
VSD	51.10 ± 16.96	50.29 ± 9.57	54.34 ± 11.89	**56.86 ± 15.34**
AVSD	53.38 ± 11.42	53.95 ± 11.43	54.69 ± 9.08	**59.26 ± 16.84**
Arrhythmia	39.62 ± 12.87	41.48 ± 9.99	41.23 ± 9.68	**43.99 ± 11.12**
Defects = 4				
Macro	64.35 ± 4.26	61.31 ± 7.14	60.60 ± 4.95	**67.63 ± 5.33**
Normal	83.38 ± 2.84	82.83 ± 2.56	82.15 ± 2.62	**84.72 ± 2.77**
VSD	45.83 ± 10.41	41.69 ± 9.37	37.74 ± 7.53	**46.70 ± 11.88**
AVSD	49.10 ± 15.37	45.42 ± 17.56	43.65 ± 18.64	**55.98 ± 17.38**
Arrhythmia	27.36 ± 4.12	**34.43 ± 9.37**	29.94 ± 5.86	31.14 ± 9.88
Cardiomegaly	**56.18 ± 11.40**	51.22 ± 13.63	51.49 ± 11.50	55.68 ± 17.09

Table 2. Multiclass classification with an increasing number of defects

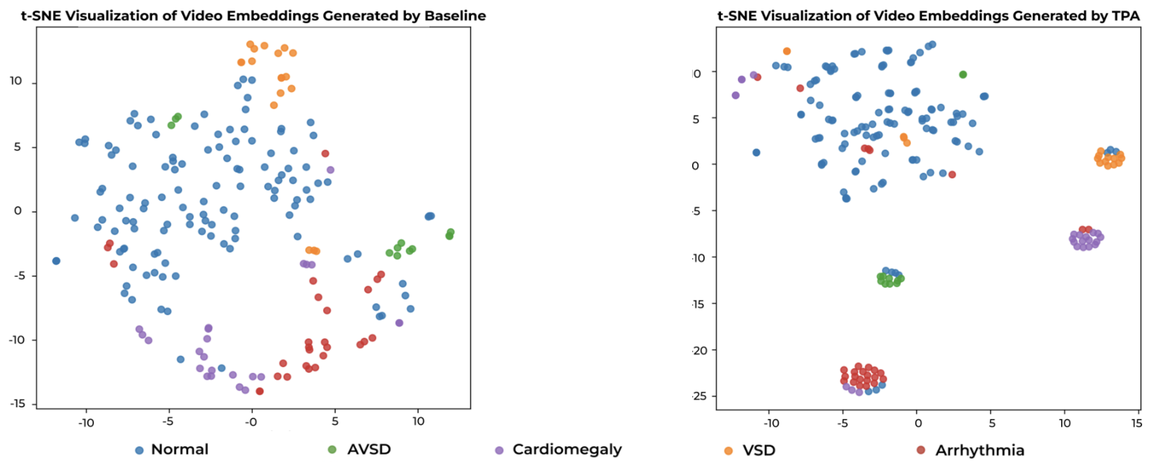


Fig 4. t-SNE visualizations of video embeddings

TAKEAWAYS

- Incorporating prompts increases video understanding and feature separation across different temporal extractors.
- Adopting TPA for a new task and an external dataset improves results when using only temporal extractors, and not just foundation models.
- CVAESM improves model reliability by reducing overconfidence and lowering calibration error while maintaining accuracy.