Plane-Wise Retrieval Memory and Structural Evidence Fusion for Tetralogy of Fallot Video Diagnosis

Xingguo Lv[1†], Kai Xu[1†], Bocheng Liang[2], Zilang Qiu[3], Jintao Wu[4], Lei Zhao[1], Guannan He[5], Chaoning Huang[6], Yuanhao Liang[7], Pengchen Liang[8], Bin Pu[1✉], and Kenli Li[1✉]

[1] Hunan University, Changsha, China
{pubin, lkl}@hnu.edu.cn
[2] Shenzhen Maternity and Child Healthcare Hospital, Shenzhen, China
[3] Beijing Normal University at Zhuhai, Zhuhai, China
[4] Sun Yat-sen University, Guangzhou, China
[5] Sichuan Provincial Maternity & Child Health Care Hospital, Chengdu, China
[6] Maternal and Child Health Hospital of Guigang City, Guigang, China
[7] Dongguan Maternal and Child Health Hospital, Dongguan, China
[8] Neusoft Medical Systems Co., Ltd, Shanghai, China

Abstract. Fetal echocardiography screening videos are acquired under continuous probe motion, so imaging planes change over time and diagnostically decisive cues for Tetralogy of Fallot (TOF) may appear only transiently and only in specific views. Conventional video classifiers treat ultrasound clips as generic temporal signals, which often misses flash-by findings and fails under incomplete plane coverage and cross-hospital domain shift. We propose a plane-aware VLM framework for TOF screening that explicitly exploits view evolution as a key variable. First, plane-labeled frames are treated as keyframes and expanded with local neighborhoods to capture short-lived anatomical cues. Second, a vision–language model performs plane-aware checklist reasoning: TOF criteria are decomposed into view-appropriate, visually checkable questions, while cues that are not assessable in the current plane are marked as such to reduce over-interpretation. Third, a plane-wise memory bank stores normal/TOF prototypes and provides retrieval-based evidence via similarity log-likelihood ratios, enabling robust inference when plane labels are missing and improving generalization across hospitals. Case-level diagnosis fuses retrieval evidence with aggregated checklist cues, with plane-coverage modulation to upweight informative but under-represented views. Experiments on multi-center fetal echo screening videos show that the proposed method improves sensitivity and cross-site robustness compared with standard video classification and VLM baselines, while producing concise, auditable findings aligned with clinical screening practice.

Keywords: Plane-aware fetal echocardiography · Tetralogy of Fallot screening · Medical Multimodal VLMs.

† Equal contribution. ✉ Corresponding authors: Bin Pu and Kenli Li.

1 Introduction

Tetralogy of Fallot (TOF) is a major conotruncal congenital heart defect and a leading cause of cyanotic congenital heart disease, with an estimated incidence on the order of a few per 10,000 live births [1, 4]. Prenatal identification is clinically important because it enables timely referral, parental counseling, delivery planning, and coordinated postnatal management. In routine obstetric screening, fetal cardiac assessment is not a single still image but a dynamic acquisition process under **continuous probe motion**, where standard planes are obtained sequentially. Practice guidelines emphasize that screening should extend beyond the four-chamber view to include outflow-tract and great-vessel views, as this substantially improves detection of major cardiac malformations, including TOF [5, 8, 20].

Despite these recommendations, reliable TOF detection in real-world screening videos remains challenging. First, diagnostic cues can be transient and plane-dependent. For example, aortic override is primarily assessed in the left ventricular outflow tract (LVOT), RVOT obstruction and pulmonary stenosis are most informative in the right ventricular outflow tract (RVOT)/three-ventricle tract (3VT), and ventricular septal defect (VSD) is commonly evaluated in four- and five-chamber views. Second, plane coverage is often incomplete in practice due to operator variability, fetal position, acoustic shadowing, and limited time, producing videos that contain only a subset of recommended views. Third, multi-center deployment introduces **domain shifts** (scanner vendors, protocols, sonographer habits), so models trained on a single hospital may degrade when evaluated externally, which is a key barrier to clinical translation [5, 17, 19, 23].

Deep learning has achieved strong performance in echocardiographic plane recognition and anatomy-centric analysis, supporting automated pipelines for clinical workflows [15, 24, 30]. However, many existing approaches for fetal cardiac screening still rely on frame-based classification or generic clip-based video modeling, where the video is treated as a homogeneous temporal signal [11, 16, 25]. Such designs are mismatched to screening: (i) short-lived findings may be missed by uniform sampling, (ii) evidence from different planes is pooled without explicit plane-aware reasoning, and (iii) models can become brittle when the scan lacks certain views or when evaluated across institutions. Although video-based detection and segmentation has been explored, robust diagnosis of multi-cue defects like TOF still requires explicit handling of view evolution and incomplete plane coverage [17].

Vision–language models (VLMs) offer an appealing alternative because they can be instructed to follow clinical protocols and produce concise, auditable intermediate findings [18, 21, 22]. Yet medical VLMs [6, 12, 26] are also known to hallucinate or over-interpret under uncertainty, especially when the visual evidence is weak or the query implicitly invites speculation, which is problematic for safety-critical screening [2, 27, 28]. This motivates a design that uses language not as free-form explanation, but as a structured interface that (i) routes questions to the appropriate plane, (ii) separates "not assessable" from "not present", and (iii) complements the VLM with retrieval evidence for cross-site robustness.

This paper presents a plane-aware VLM framework for TOF screening in fetal echocardiography videos, explicitly leveraging view evolution metadata to handle transient, view-dependent cues. (1) **Plane-labeled frames are used as keyframes and expanded with local neighborhoods**, increasing the probability of capturing transient cues without densifying computation; (2) We formulate TOF diagnosis as a compact **plane-aware checklist reasoning protocol**: each hallmark is queried only in clinically assessable planes, and non-assessable cues are explicitly labeled to reduce over-interpretation under partial visibility; (3) A plane-wise memory bank stores normal/TOF prototypes and provides **retrieval-based evidence**, enabling inference when plane labels are missing and improving cross-hospital generalization; (4) We curate a **multi-center TOF screening video dataset** and conduct extensive comparisons and cross-domain evaluations across hospitals, demonstrating improved sensitivity and robustness over standard video classification and VLM baselines.

2 Method

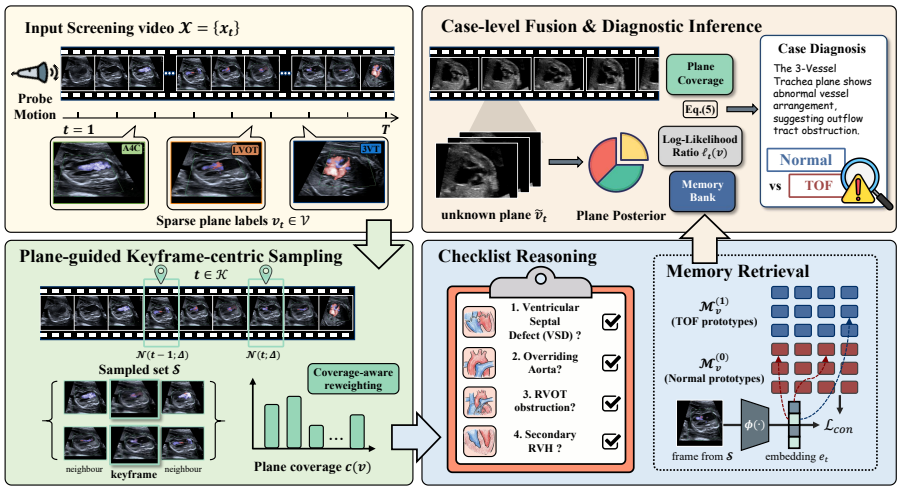

Fig. 1. Pipeline of the proposed plane-aware TOF screening framework. Keyframe-centric sampling leverages sparse plane labels to gather neighborhood frames and estimate plane coverage. Checklist reasoning, implemented via a VLM, performs plane-aware assessment of TOF hallmarks, while plane-wise memory retrieval provides normal/TOF matching evidence. At inference, case-level fusion integrates these signals to produce the final TOF diagnosis and a concise case summary.

2.1 Problem Setup and Notation

A fetal cardiac screening video is represented as a frame sequence $\mathcal{X} = \{x_t\}_{t=1}^{T}$, where $x_t \in \mathbb{R}^{H \times W}$ denotes the t-th ultrasound frame. During clinical scanning, the sonographer continuously moves the probe, causing the imaging plane to

change over time. Accordingly, a subset of frames are annotated with plane labels $v_t \in \mathcal{V}$ (e.g., A4C, LVOT, 3VT), where $\mathcal{V}$ is a finite set of standard planes. The case-level diagnosis label is binary: $y \in \{0, 1\}$, where $y = 1$ indicates Tetralogy of Fallot (TOF).

The goal is to predict $\hat{y}$ from $\mathcal{X}$ by explicitly leveraging the plane evolution and the fact that TOF-relevant anatomical cues may be transient and plane-dependent. Figure 1 illustrates the pipeline: (i) plane-guided keyframe-centric sampling, (ii) plane-aware checklist reasoning with a VLM, and (iii) plane-wise memory retrieval for missing or incomplete plane labels.

2.2 Plane-guided Keyframe-centric Sampling

Conventional video classification treats $\mathcal{X}$ as a generic temporal signal and typically uses uniform sampling or fixed-length clips [11, 16, 25]. In fetal echo screening, however, clinically decisive structures (e.g., a membranous VSD or overriding aorta) may only appear for a few frames and often occur outside a single canonical plane. Therefore, plane changes provide a natural "state" variable to organize sampling.

Let $\mathcal{K} = \{t \in \{1, \ldots, T\} \mid v_t\}$ denote indices of frames with plane labels. These frames are treated as **keyframes**. For each keyframe $t \in \mathcal{K}$, a local neighborhood is sampled:

$$\mathcal{N}(t; \Delta) = \{t - \Delta, \ldots, t + \Delta\} \cap \{1, \ldots, T\}, \tag{1}$$

where Δ controls the temporal window. The final sampled set is

$$\mathcal{S} = \bigcup_{t \in \mathcal{K}} \mathcal{N}(t; \Delta). \tag{2}$$

This keyframe-centric strategy ensures that when the plane label indicates a diagnostically relevant moment, nearby frames are also included to capture short-lived cues and reduce sensitivity to a single noisy frame.

To mitigate cases where the scan contains only partial planes (e.g., only A4C), a plane-coverage vector $\mathbf{c}$ is computed:

$$\mathbf{c} \in \mathbb{R}^{|\mathcal{V}|}, \quad c(v) = \frac{1}{|\mathcal{K}|} \sum_{t \in \mathcal{K}} \mathbb{I}[v_t = v]. \tag{3}$$

$\mathbf{c}$ later modulates fusion weights (Sec. 2.4) so that evidence from under-represented but highly informative planes can be upweighted if present.

2.3 Checklist Reasoning from Plane-labeled Keyframes

TOF diagnosis in screening videos is inherently *multi-cue and plane-dependent*: the four hallmark components, denoted by F (VSD, overriding aorta, RVOT obstruction, and secondary RVH) are not simultaneously observable in a single frame or a single standard plane, and some cues may only appear transiently

during probe motion. Instead of asking the VLM to directly output a label, we convert these medical criteria into a compact *plane-aware screening protocol*. Concretely, each TOF component is decomposed into a small set of visually checkable questions (e.g., septal discontinuity consistent with VSD; aortic root alignment/override relative to septum; RVOT caliber/flow-tract narrowing; RV wall thickening as a secondary sign), and each question is only activated under planes where it is clinically assessable (e.g., VSD/override mainly in A4C/LVOT, RVOT obstruction mainly in RVOT/3VT). The provided view-context acts as a routing signal that restricts which checks are considered reliable at each moment: when the current plane does not support a cue, the model is instructed to treat it as not assessable rather than speculating, which reduces hallucination under partial visibility and noisy frames. As the probe moves and planes change, complementary checks become available; the downstream fusion stage aggregates these plane-consistent findings across the sampled keyframe neighborhoods to reach a case-level decision.

2.4 Memory-augmented Retrieval and Case-level Fusion

To improve cross-hospital generalization and to support videos with incomplete plane annotations, we introduce a plane-wise memory and fuse retrieval evidence with plane-consistent checklist findings. A visual encoder $\phi(\cdot)$ maps each sampled frame to an embedding $e_t = \phi(x_t) \in \mathbb{R}^d$. For each plane $v \in \mathcal{V}$ and class $y \in \{0, 1\}$, a memory bank stores prototypes $\mathcal{M}_v^{(y)} = \{m_{v,k}^{(y)}\}_{k=1}^{K_y}$. The memory bank is constructed exclusively from training and remains fixed during inference. This design prevents information leakage while providing stable plane-conditioned reference patterns for retrieval. Given a plane label $\tilde{v}_t$, we compute a similarity-based log-likelihood ratio (LLR) as retrieval evidence:

$$\ell_t(\tilde{v}_t) = \log \sum_k \exp\Big(\frac{\cos(e_t, m_{\tilde{v}_t,k}^{(1)})}{\tau}\Big) - \log \sum_k \exp\Big(\frac{\cos(e_t, m_{\tilde{v}_t,k}^{(0)})}{\tau}\Big). \tag{4}$$

If plane labels are missing, we first infer a plane posterior from the same memory (view recognition by prototype similarity) and marginalize ℓ_t across planes; this keeps retrieval usable in general screening clips without reliable plane metadata.

Case-level diagnosis fuses retrieval evidence and checklist cues over the sampled set $\mathcal{S}$:

$$L = \sum_{t \in \mathcal{S}} \omega(\tilde{v}_t)\, \tilde{\ell}_t + \sum_{f \in F} g_f, \quad \hat{y} = \sigma(L), \tag{5}$$

where $\omega(v) = \frac{1}{c(v)+\epsilon}$ is a reliability-aware weight based on plane coverage of the video in Eq. (3), and $\tilde{\ell}_t$ equals ℓ_t when the plane is known (otherwise the plane-marginalized LLR). For checklist fusion, g_f is computed by a noisy-or style aggregation across frames that are compatible with cue f, reflecting the screening logic that *any* strong supporting observation can trigger a positive decision while non-assessable cues remain neutral.

Table 1. Statistics of the multi-center fetal echocardiogram datasets. The external centers exhibit substantial heterogeneity in both disease prevalence and temporal acquisition attributes.

Dataset	Total Videos	Normal (%)	Abnormal (%)	Duration (s)	Avg. FPS
Internal	893	500 (56.0%)	393 (44.0%)	7.66 ± 2.1	44.70
External 1	54	26 (48.1%)	28 (51.9%)	16.62 ± 4.2	25.26
External 2	105	51 (48.6%)	54 (51.4%)	7.77 ± 1.8	45.21
External 3	103	50 (48.6%)	53 (51.4%)	12.64 ± 3.5	69.74
Total	**1,155**	**627**	**528**	-	-

Training uses patient-level binary cross-entropy on $\hat{y}$, together with a plane-aware contrastive term to shape embeddings for retrieval:

$$\mathcal{L}_{\text{con}} = -\log \frac{\sum_k \exp(\cos(e_t, m^{(y)}_{\tilde{v}_t,k})/\tau)}{\sum_{y' \in \{0,1\}} \sum_k \exp(\cos(e_t, m^{(y')}_{\tilde{v}_t,k})/\tau)}. \tag{6}$$

At inference, we apply keyframe-centric sampling when plane labels exist; otherwise, we sample periodically and marginalize retrieval evidence using the inferred plane posterior.

3 Experiments

3.1 Datasets and Experimental Setup

Multi-center Cohort. We curated 1,155 fetal echocardiogram videos from an international fetal birth-defect research consortium spanning four geographically distinct regions. (1) **Internal Dataset:** 893 subjects (56.0% normal) from Hospital 1 for development. (2) **External Validation:** Three sets for zero-shot evaluation: **External 1** ($n = 54$, 25.3 FPS), **External 2** ($n = 105$, 45.2 FPS), and **External 3** ($n = 103$, 69.7 FPS), all with ∼51% abnormality. Table 1 details parameters and distributions. All procedures were approved by the local ethics committee (LLYJ2024-370-121).

Implementation. Built on Qwen3-VL-8B-Instruct [3], videos were resized to 256^2 and sampled at 4 FPS ($T = 32$). The visual encoder is optimized by Eq. (6) and construct the prototype memory. We used LoRA ($r = 32, \alpha = 64$) on 8 NVIDIA RTX A6000 GPUs, freezing the vision tower and LLM backbone. Training employed AdamW (LR 2×10^{-4}) with a cosine scheduler over 5 epochs (batch size 64).

Baselines and Metrics. We compared against: (1) **Conventional Video Classifiers**: UniFormerV2 [14], Video-MAEv2 [25]; (2) **General-purpose Multimodal VLMs**: Qwen3-VL-8B [3], GPT-4o [9]; (3) **Medical Multimodal VLMs**: HuatuoGPT-7B [6], LLaVA-Med-7B [12], Health-GPT-M3 [18], Hulu-Med-7B [10]; (4) **Video-based VLMs**: MiniCPM-V [29], VideoChat2-HD [13], VideoLLaMA2 [7], and F-16-7B [16]. Metrics include F1 (primary), Sensitivity (Sen), Specificity (Spe), AUC, Precision (Pre), and Accuracy (Acc).

Table 2. Performance on internal and three external sets.

Method	Internal						External 1						External 2						External 3					
	F1	Sen	Spe	AUC	Pre	Acc	F1	Sen	Spe	AUC	Pre	Acc	F1	Sen	Spe	AUC	Pre	Acc	F1	Sen	Spe	AUC	Pre	Acc
Conventional Video Classifiers																								
UniFormerV2 [14]	69.1	70.0	72.9	71.5	68.3	71.6	61.8	60.7	61.5	61.1	63.0	61.1	61.1	61.1	58.8	59.9	61.1	60.0	55.2	54.7	54.0	54.3	55.8	54.4
Video-MAEv2 [25]	72.3	75.0	72.9	74.0	69.8	73.9	64.3	64.3	61.5	62.9	64.3	63.0	65.5	66.7	60.8	63.7	64.3	63.8	58.5	58.5	56.0	57.2	58.5	57.3
General-purpose Multimodal VLMs																								
Qwen3-VL-8B [3]	75.8	76.7	76.6	74.9	75.0	76.7	48.7	35.7	50.0	41.5	76.9	38.2	62.5	59.5	43.1	50.9	65.8	53.7	16.6	30.7	38.0	32.1	11.4	36.5
ChatGPT-4o [9]	70.8	57.5	95.8	76.7	92.0	78.4	55.0	39.3	96.2	67.7	91.7	66.7	62.1	50.0	88.2	69.1	81.8	68.6	54.8	43.4	84.0	63.7	74.2	63.1
Video-based VLMs																								
MiniCPM-V [29]	44.7	47.5	45.8	46.7	42.2	46.6	42.9	42.9	38.5	40.7	42.9	40.7	40.7	40.7	37.3	39.0	40.7	39.0	38.1	37.7	36.0	36.9	38.5	36.9
VideoChat2-HD [13]	48.8	52.5	47.9	50.2	45.7	50.0	46.4	46.4	42.3	44.4	46.4	44.4	42.6	42.6	39.2	40.9	42.6	41.0	40.0	39.6	38.0	38.8	40.4	38.8
VideoLLaMA2-7B [7]	70.6	60.0	91.7	75.8	85.7	77.3	55.0	39.3	96.2	67.7	91.7	66.7	62.9	51.9	86.3	69.1	80.0	68.6	55.8	45.3	82.0	63.6	72.7	63.1
F-16-7B [16]	68.0	87.5	41.7	64.6	55.6	62.5	65.7	78.6	34.6	56.6	56.4	57.4	64.7	83.3	21.6	52.5	52.9	53.3	63.1	77.4	28.0	52.7	53.2	53.4
Medical Multimodal VLMs																								
HuatuoGPT-7B [6]	65.1	67.5	66.7	67.1	62.8	67.0	57.1	57.1	53.8	55.5	57.1	55.6	58.7	59.3	52.9	56.1	57.1	56.2	52.8	52.8	48.0	50.4	51.9	50.5
LLaVA-Med-7B [12]	69.0	72.5	70.8	71.7	67.4	71.6	61.8	60.7	61.5	61.1	63.0	61.1	63.6	64.8	56.9	60.9	61.4	61.0	56.1	56.6	52.0	54.3	55.6	54.4
Health-GPT-M3 [18]	82.6	95.0	70.8	82.9	73.1	81.8	72.7	**85.7**	46.2	65.9	63.2	66.7	71.9	**92.6**	31.4	62.0	58.8	62.9	64.3	**86.8**	12.0	49.4	51.1	50.5
Hulu-Med-7B [10]	84.1	72.5	**100.0**	86.2	**100.0**	87.5	79.2	67.9	96.2	82.0	95.0	81.5	73.3	61.1	**94.1**	77.6	**91.7**	77.1	60.0	45.3	94.0	69.6	88.9	68.9
Ours	**95.1**	**97.5**	93.7	**95.6**	92.8	**95.4**	**88.2**	79.7	**98.0**	**88.9**	**98.6**	**88.5**	**75.0**	92.3	86.0	**89.1**	63.1	**89.2**	**69.7**	53.5	**100.0**	**76.7**	**100.0**	**76.1**

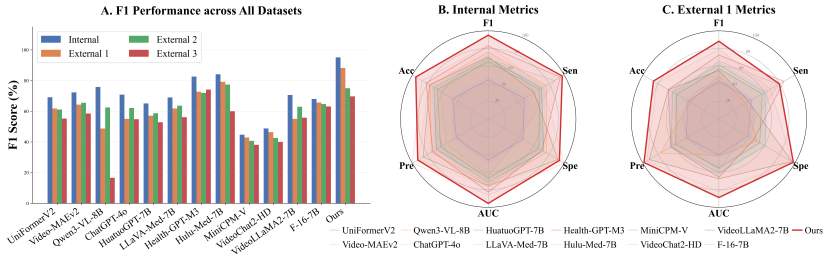

Fig. 2. Performance comparison against SOTA baselines. (A) F1 scores across cohorts. (B, C) Radar charts of multi-metric evaluation on Internal and External 1 datasets.

3.2 Comparative Analysis and Visualization

As shown in Table 2 and Fig. 2, our framework consistently outperforms all baselines.

Performance and Robustness. Conventional video backbones [14, 25] and Video-VLMs [7, 13, 29] yield sub-optimal results (best internal F1 72.3%) by overlooking subtle plane transitions critical for TOF diagnosis [4]. While medical MLLMs [10, 18] improve internal accuracy, they exhibit severe sensitivity-specificity imbalance. General models [3, 9] show erratic behavior, with Qwen3-VL's F1 collapsing to 16.6% on External 3. In contrast, our method achieves SOTA performance internally (95.1% F1) and maintains superior generalization (Ext. Avg. F1 77.7%). Critically, we preserve balanced sensitivity (97.5%) and specificity (93.7%), minimizing missed diagnoses [5].

Reasoning Trace. As visualized in Fig. 3, our model performs structured anatomical evaluation guided by clinical protocols. For normal cases, it confirms septal integrity across 4CV/LVOT views; for TOF cases, it captures transient cues like abnormal vessel arrangement in the 3VT view. These gains stem from: (i) plane-guided sampling; (ii) checklist-based reasoning that mitigates hallucinations [27] in ambiguous views; and (iii) plane-wise memory retrieval. This structured trace aligns with ISUOG guidelines, enhancing the reliability of AI-assisted screening.

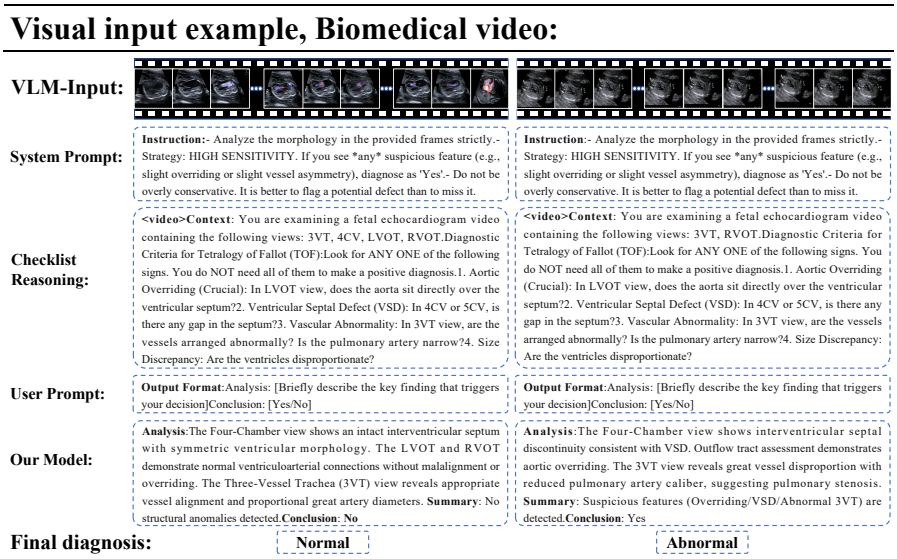

Fig. 3. Visualization of plane-aware reasoning. The model identifies key TOF features (e.g., abnormal 3VT arrangement) via a clinical checklist, providing an interpretable summary.

Table 3. Ablation study on internal and external test sets. Metrics are in %. "Ext Avg" is the mean over External 1/2/3; "Worst Ext" is the worst among them.

Variant	Internal				Ext Avg				Worst Ext			
	F1	AUC	Sen	Spe	F1	AUC	Sen	Spe	F1	AUC	Sen	Spe
Baseline	75.9	76.7	76.7	76.6	42.7	42.9	42.0	43.7	16.7	34.4	30.8	38.0
Sampling	78.9	82.6	65.1	100.0	49.7	54.9	59.9	49.9	32.7	48.8	28.6	15.7
Protocol	91.1	91.9	83.7	100.0	69.3	73.4	62.8	83.9	58.5	63.1	42.9	76.5
GlobalMem†	94.2	95.0	96.5	92.0	74.9	82.9	72.5	92.0	66.8	74.6	50.5	84.0
NoCL†	94.5	95.1	96.8	92.5	75.9	83.6	73.5	92.5	67.8	75.4	51.5	84.5
Full	95.1	95.6	97.5	93.7	77.7	85.0	75.2	94.7	69.8	76.8	53.6	86.0

3.3 Ablation Analysis

Table 3 evaluates each module under one-hospital training and three-hospital external testing. Plane-guided **Sampling** yields a modest gain but can over-flag in the hardest domain (Worst Ext Spe 38.0→15.7), whereas adding the plane-aware **Protocol** gives the largest cross-domain improvement and restores specificity (Ext Avg F1/AUC 49.7/54.9→69.3/73.4; Spe 49.9→83.9). Replacing plane-wise memory with a **GlobalMem** slightly reduces external robustness, and removing contrastive embedding shaping (**NoCL**) causes a further drop, supporting the role of plane-specific prototypes and discriminative retrieval. **Full** achieves the best overall external performance (Ext Avg F1/AUC 77.7/85.0) and improves worst-case results (Worst Ext F1 58.5→69.8).

4 Conclusion

We presented a plane-aware VLM framework for TOF screening in fetal echocardiography videos that leverages view evolution to address transient, plane-dependent cues and incomplete plane coverage. By combining plane-guided keyframe-centric sampling, plane-aware checklist reasoning, and a plane-wise prototype memory for retrieval-based evidence, the proposed approach yields a concise and auditable case-level diagnosis and improves cross-hospital robustness. Extensive multi-center experiments demonstrate consistent gains over conventional video classification and VLM baselines.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62227808, Grant 62506124, and in part by the Natural Science Foundation of Hunan Province under Grants 2025JJ60408.

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

References

1. Apitz, C., Webb, G.D., Redington, A.N.: Tetralogy of fallot. The Lancet **374**(9699), 1462–1471 (2009)
2. Artsi, Y., Klang, E., Collins, J.D., Glicksberg, B.S., Nadkarni, G.N., Korfiatis, P., Sorin, V.: Large language models in radiology reporting-a systematic review of performance, limitations, and clinical implications. Intelligence-Based Medicine p. 100287 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bailliard, F., Anderson, R.H.: Tetralogy of fallot. Orphanet journal of rare diseases **4**(1), 2 (2009)
5. Carvalho, J., Axt-Fliedner, R., Chaoui, R., Copel, J., Cuneo, B., Goff, D., Gordin Kopylov, L., Hecher, K., Lee, W., Moon-Grady, A., et al.: Isuog practice guidelines (updated): fetal cardiac screening. Ultrasound Obstet Gynecol **61**(6), 788–803 (2023)
6. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
7. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024), https://arxiv.org/abs/2406.07476
8. Donofrio, M.T., Moon-Grady, A.J., Hornberger, L.K., Copel, J.A., Sklansky, M.S., Abuhamad, A., Cuneo, B.F., Huhta, J.C., Jonas, R.A., Krishnan, A., et al.: Diagnosis and treatment of fetal cardiac disease: a scientific statement from the american heart association. Circulation **129**(21), 2183–2242 (2014)
9. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

10. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
11. KunChang Li, Yinan He, Y.W.Y.L.W.W.P.L.Y.W.L.W., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS (2023)
13. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
14. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Unlocking the potential of image vits for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1632–1643 (2023)
15. Li, X., Zhang, H., Yue, J., Yin, L., Li, W., Ding, G., Peng, B., Xie, S.: A multi-task deep learning approach for real-time view classification and quality assessment of echocardiographic images. Scientific Reports **14**(1), 20484 (2024)
16. Li, Y., Tang, C., Zhuang, J., Yang, Y., Sun, G., Li, W., Ma, Z., Zhang, C.: Improving llm video understanding with 16 frames per second. In: International Conference on Machine Learning. pp. 35942–35956. PMLR (2025)
17. Liastuti, L.D., Nursakina, Y.: Diagnostic accuracy of artificial intelligence models in detecting congenital heart disease in the second-trimester fetus through prenatal cardiac screening: a systematic review and meta-analysis. Frontiers in Cardiovascular Medicine **12**, 1473544 (2025)
18. Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., Song, X., et al.: Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. arXiv preprint arXiv:2502.09838 (2025)
19. Lv, X., Dong, X., Wang, L., Yang, J., Zhao, L., Pu, B., Jin, Z., Li, X.: Test-time domain generalization via universe learning: A multi-graph matching approach for medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15621–15631 (2025)
20. Moon-Grady, A.J., Donofrio, M.T., Gelehrter, S., Hornberger, L., Kreeger, J., Lee, W., Michelfelder, E., Morris, S.A., Peyvandi, S., Pinto, N.M., et al.: Guidelines and recommendations for performance of the fetal echocardiogram: an update from the american society of echocardiography. Journal of the American Society of Echocardiography **36**(7), 679–723 (2023)
21. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14788–14798 (2025)
22. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
23. Pu, B., Lv, X., Yang, J., Guannan, H., Dong, X., Lin, Y., Shengli, L., Ying, T., Fei, L., Chen, M., et al.: Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In: Forty-first International Conference on Machine Learning (2024)

24. Pu, B., Yang, J., Lv, X., Xu, K., Li, K.: Unified mixture-of-experts framework for joint cardiac and vascular ultrasound analysis and report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 8439–8447 (2026)
25. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14549–14560 (2023)
26. Wang, Z., et al.: Medclip: Contrastive learning from unpaired medical images and text. In: EMNLP (2022)
27. Wu, J., Kim, Y., Wu, H.: Hallucination benchmark in medical visual question answering. arXiv preprint arXiv:2401.05827 (2024)
28. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
29. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
30. Zhao, L., Lv, X., Lin, Q., Shi, K., Qi, X., Pu, B., Li, K.: Concept relationship embedding-based interactive web application for explainable medical diagnosis. In: Proceedings of the ACM Web Conference 2026. pp. 8973–8983 (2026)