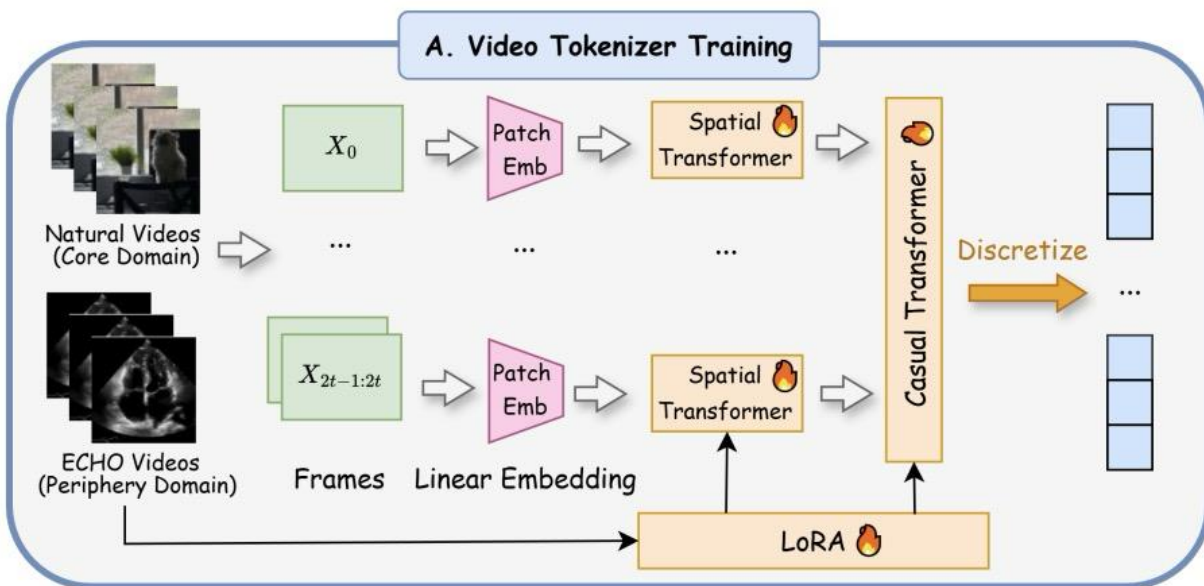


ECHOPULSE: ECG CONTROLLED ECHOCARDIOGRAMS VIDEO GENERATION

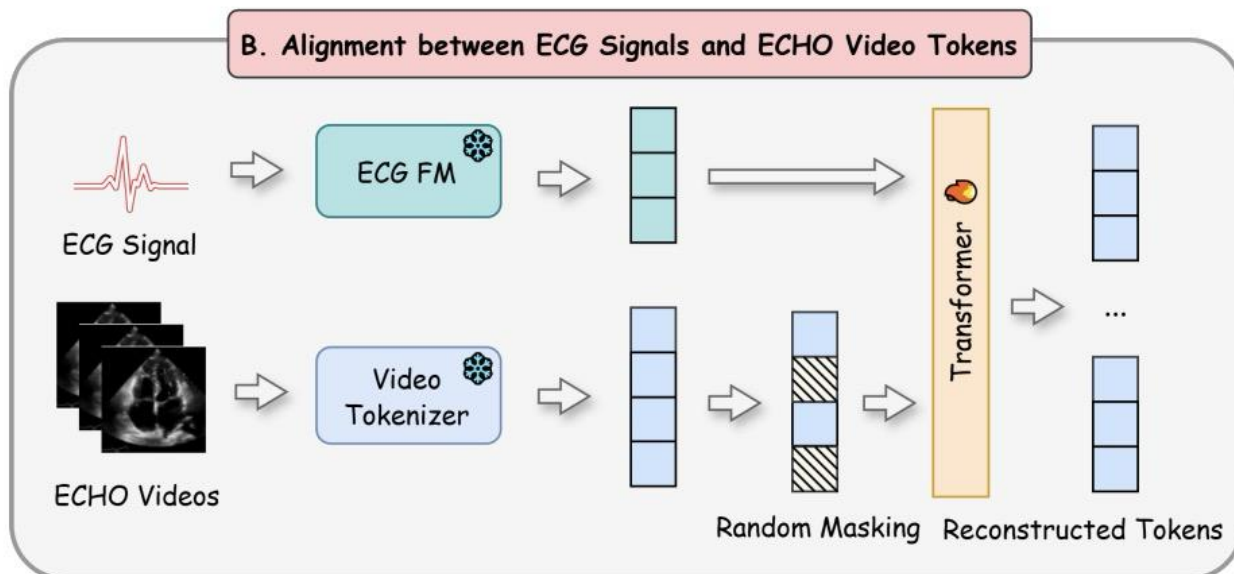
Yiwei Li^{1,2*} , Sekeun Kim^{2*} , Zihao Wu^{1,2} , Hanqi Jiang¹ , Yi Pan¹ Pengfei Jin² ,
Sifan Song² , Yucheng Shi¹ , Tianming Liu^{1†} , Quanzheng Li^{2†} , Xiang Li^{2†}

1 University of Georgia 2Massachusetts General Hospital and Harvard Medical
School

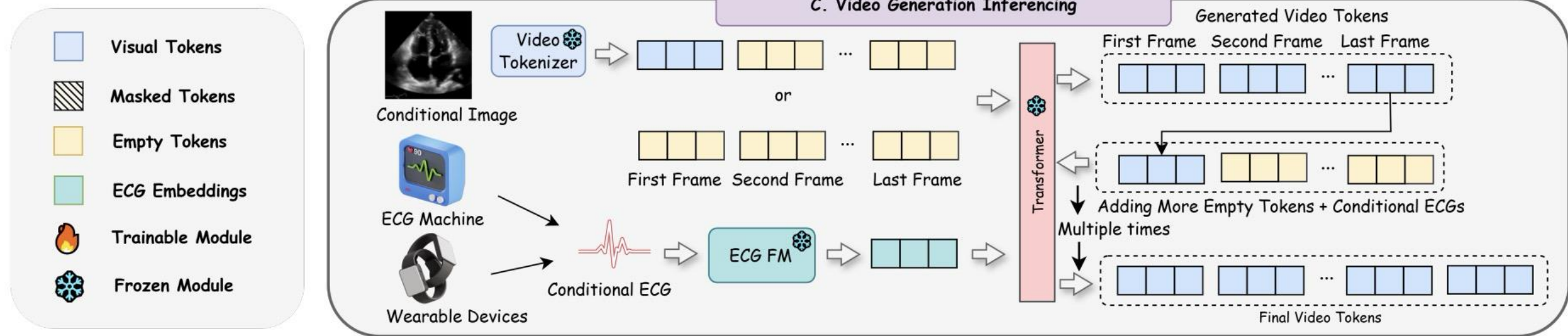
A. Video Tokenizer Training



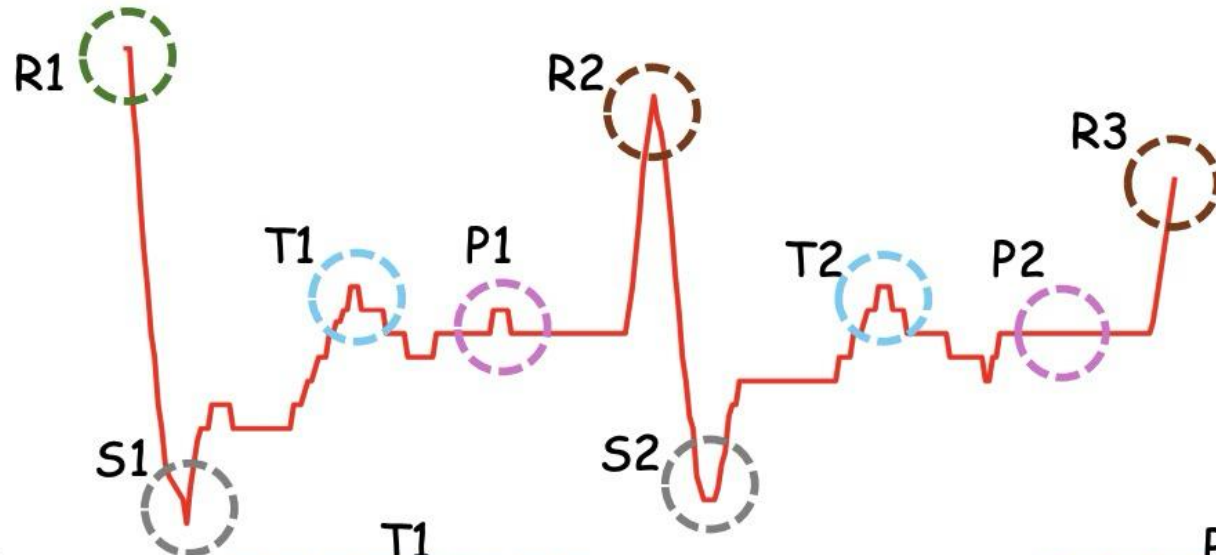
B. Alignment between ECG Signals and ECHO Video Tokens



C. Video Generation Inferencing

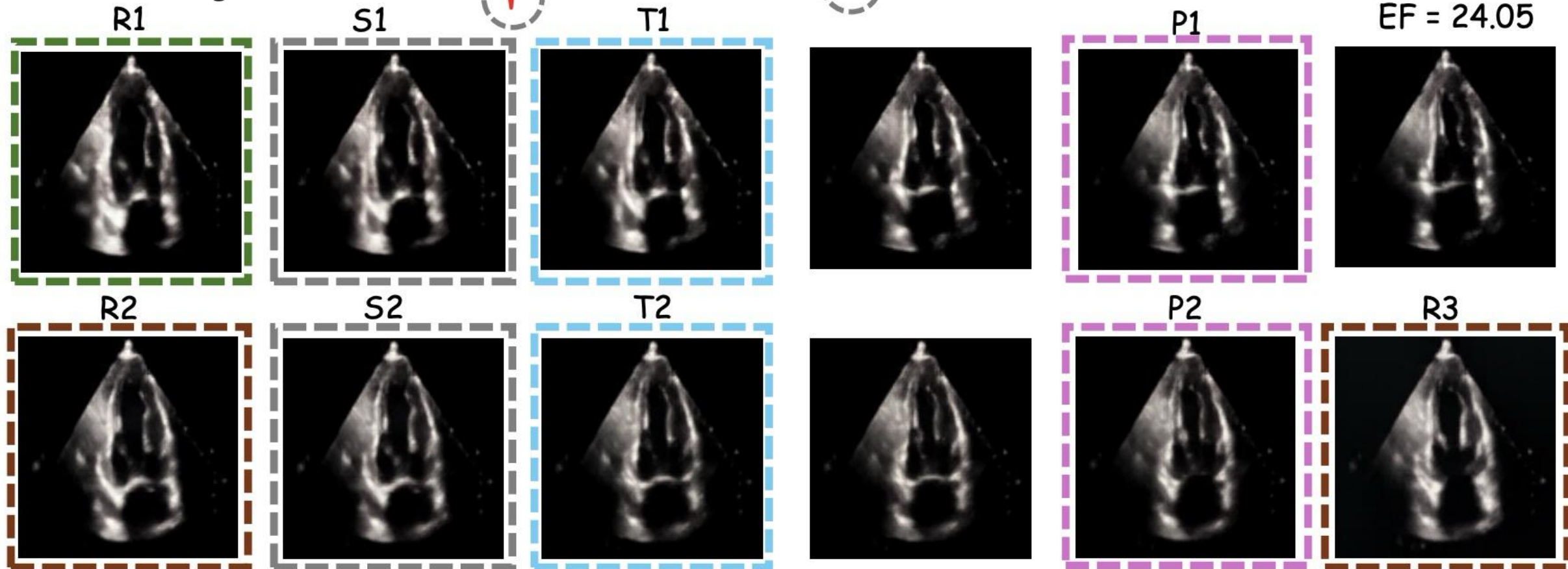


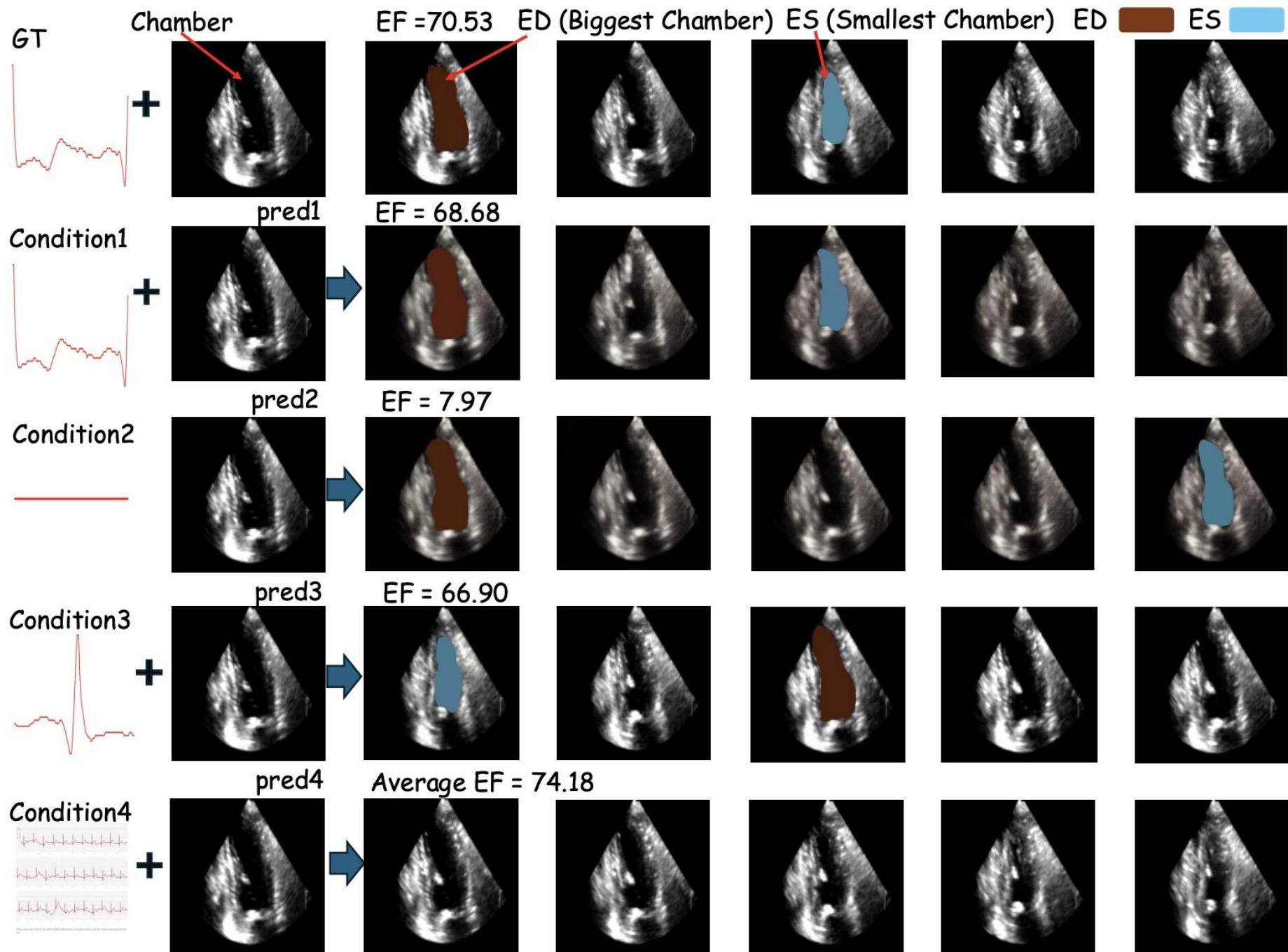
Condition ECG Signal



R : R Wave (ED)
T : T Wave (ES)
P : P Wave
S : S Wave

Condition Image





ECG collected from Apple Watch

Table 1: **Reconstruction metrics:** Videos reconstructed using video tokenization and diffusion-based models are evaluated across two datasets with MAE and MSE. The **best** results are highlighted.

Methods	Dataset	MSE↓	MAE↓
EchoNet-Synthetic (Reynaud et al., 2024)	Echo-Dynamic	3.00×10^{-3}	3.00×10^{-2}
ECHOPulse (Only natural videos)	Echo-Dynamic	5.02×10^{-3}	4.08×10^{-2}
ECHOPulse (Domain transfer)	Echo-Dynamic	2.89×10^{-3}	2.92×10^{-2}
ECHOPulse (Only natural videos)	Private data	6.09×10^{-3}	4.41×10^{-2}
ECHOPulse (Domain transfer)	Private data	2.91×10^{-3}	2.97×10^{-2}

Table 2: Image quality metrics: The videos generated by different methods under various conditions across three datasets are evaluated using three image quality metrics. The **best** and **second** results are highlighted. Note that EchoNet-Synthetic (Reynaud et al., 2024) did not report the SSIM metric and did not specify whether the results correspond to A2C or A4C views.

Methods	Dataset	Condition	A2C			A4C		
			FID↓	FVD↓	SSIM↑	FID↓	FVD↓	SSIM↑
MoonShot (Zhang et al., 2024)	CAMUS	Text	48.44	202.41	0.63	61.57	290.08	0.62
VideoComposer (Wang et al., 2024)	CAMUS	Text	37.68	164.96	0.60	35.04	180.32	0.61
HeartBeat (Zhou et al., 2024)	CAMUS	Text	107.66	305.12	0.53	76.46	381.28	0.53
ECHOPulse (Only natural videos)	CAMUS	Text	12.71	273.15	0.61	15.38	336.04	0.58
ECHOPulse (Domain transfer)	CAMUS	Text	5.65	211.85	0.79	8.17	283.32	0.75
EchoDiffusion (Reynaud et al., 2023)	Echo-Dynamic	Text+EF	24.00	228.00	0.48	24.00	228.00	0.48
ECHOPulse (Only natural videos)	Echo-Dynamic	Text	36.10	319.25	0.39	44.21	334.95	0.35
ECHOPulse (Domain transfer)	Echo-Dynamic	Text	27.50	249.46	0.46	29.83	312.31	0.41
ECHOPulse (Only natural videos)	Private data	Text	27.49	291.67	0.53	34.13	374.92	0.51
ECHOPulse(Domain transfer)	Private data	Text	25.44	224.90	0.54	31.21	334.09	0.54
ECHOPulse (Only natural videos)	Private data	ECG	18.73	200.45	0.56	27.37	302.89	0.55
ECHOPulse (Domain transfer)	Private data	ECG	15.50	82.44	0.67	20.82	107.40	0.66
HeartBeat (Zhou et al., 2024)	CAMUS	Text+Sketch+Mask+...	25.23	97.28	0.66	31.99	159.36	0.65
EchoNet-Synthetic (Reynaud et al., 2024)	Echo-Dynamic	Video+EF	16.90	87.40	-	16.90	87.40	-

Table 3: **Clinical and time-inference metrics:** This table presents the EF (Ejection Fraction) differences between the generated videos and the target ones. The sampling time refers to the time required to generate 64 video frames. The **best** results are highlighted.

Methods	Condition	$R^2 \uparrow$	MAE $\downarrow$	RMSE $\downarrow$	S. time $\downarrow$	Parameters
EchoDiffusion (Reynaud et al., 2023)	Image+EF	0.89	4.81	6.69	Gen. 146s	381M
ECHOPulse	Image+ECG	0.85	2.51	2.86	Gen. 6.4s	279M

Table 6: **Video reconstruction metrics:** This table presents the (reconstruction FID) rFID and (reconstruction LVEF) rLVEF differences between the reconstructed videos and the target ones.

Methods	rFID $\downarrow$	rLVEF
EchoNet-Synthetic(Reynaud et al., 2024)	10.93	1 ± 0.33
ECHOPulse (Our Model)	9.16	1 ± 0.26

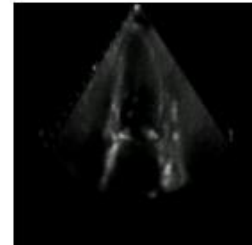
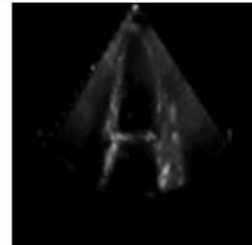
Table 7: **Prediction errors for ED and ES frames**, calculated by comparing the segmented frames from the ground-truth video and the generated video. These errors validate the consistency of cardiac motion in the generated video and provide insights into whether the model accurately follows ECG-driven dynamics.

Methods	Average Error (ED) $\downarrow$	Average Error (ES) $\downarrow$
EchoDiffusion (Reynaud et al., 2023)	1.55	1.25
EchoNet-Synthetic(Reynaud et al., 2024)	1.06	0.97
ECHOPulse (Our Model)	0.18	0.21

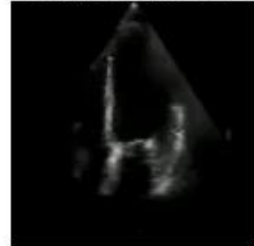
Condition Image



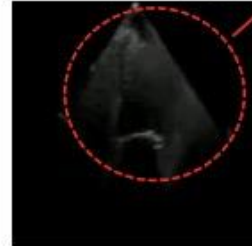
Ground-Truth



EchoDiffusion



Loss of details

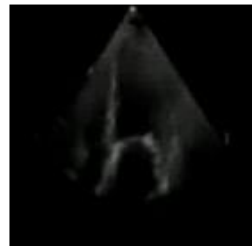


Blur caused by error accumulation



EF = 50

EchoNet-Synthetic



Valvular structural error

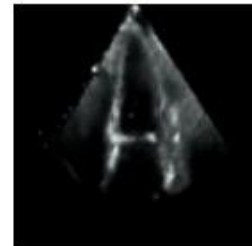
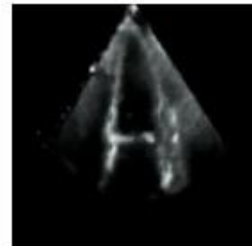
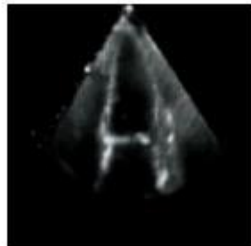
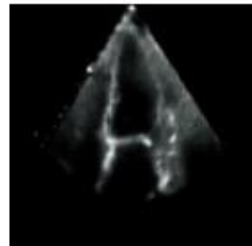
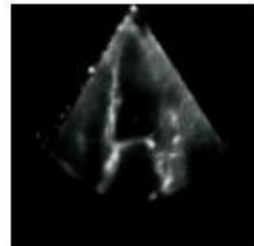


Detail blur

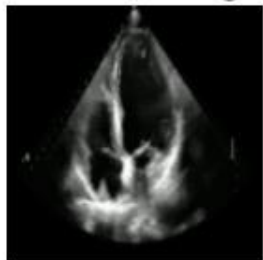


OR

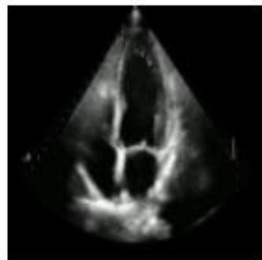
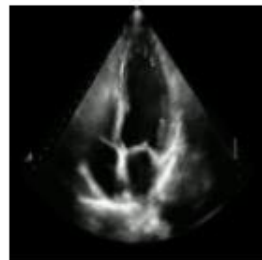
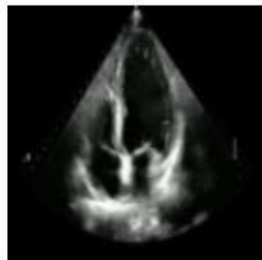
ECHOPulse



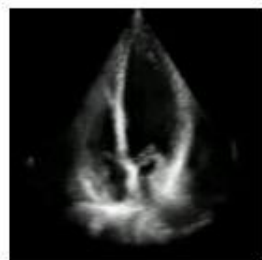
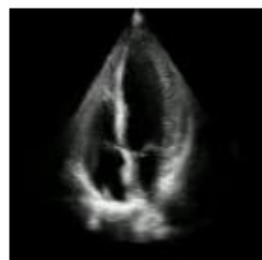
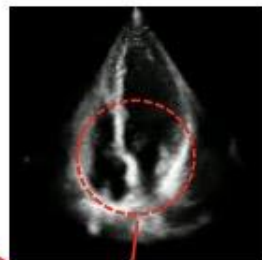
Condition Image



Ground-Truth

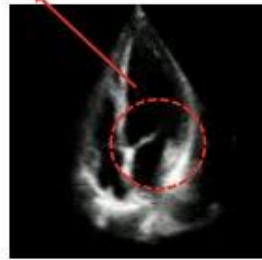
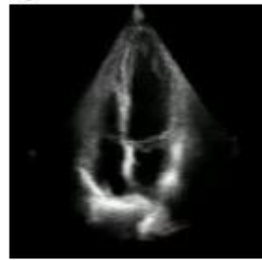
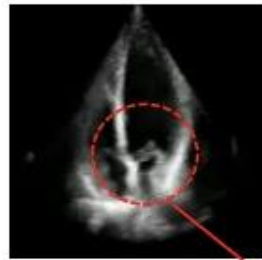
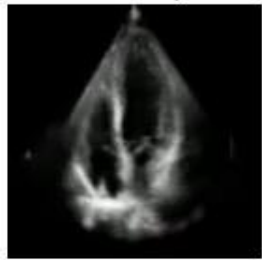


EchoDiffusion



EF = 55

EchoNet-Synthetic

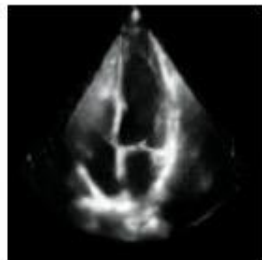
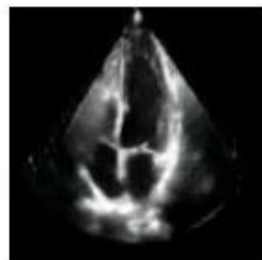
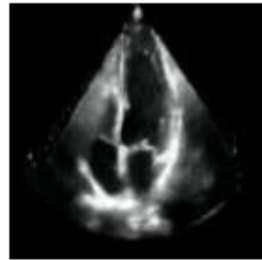
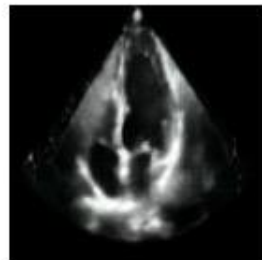
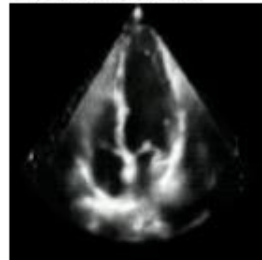


Valvular missing

Valvular motion corresponding phase error

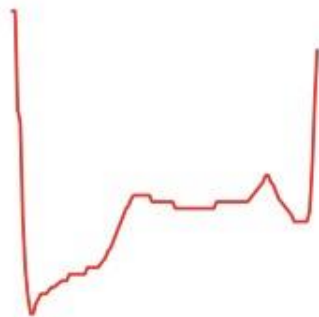
OR

ECHOPulse



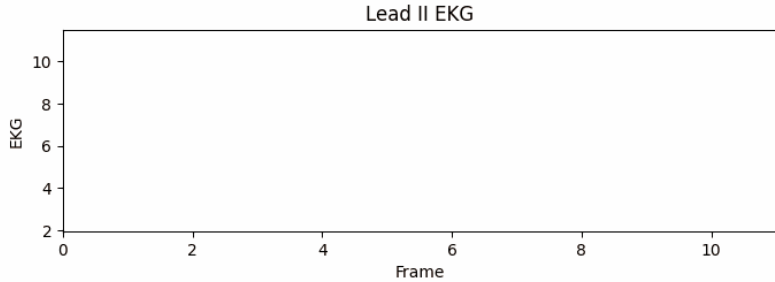
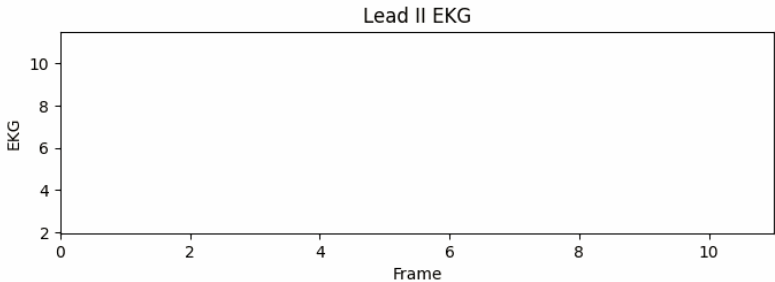
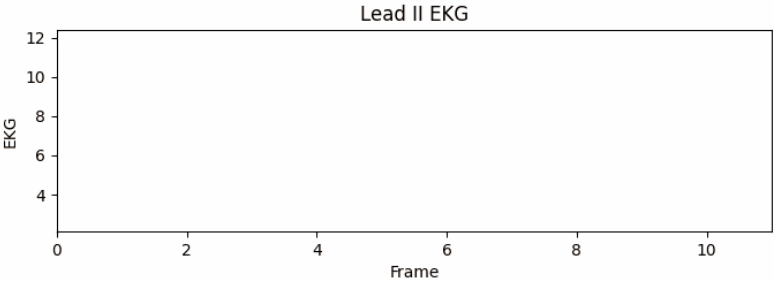
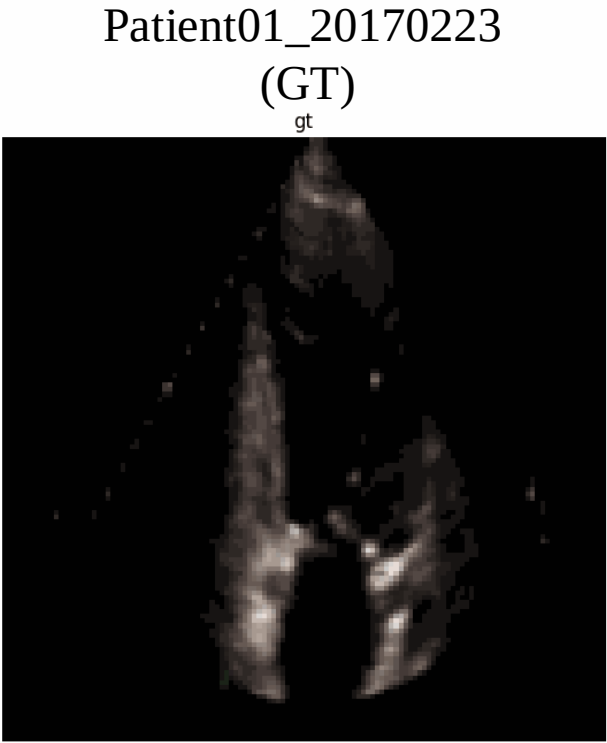
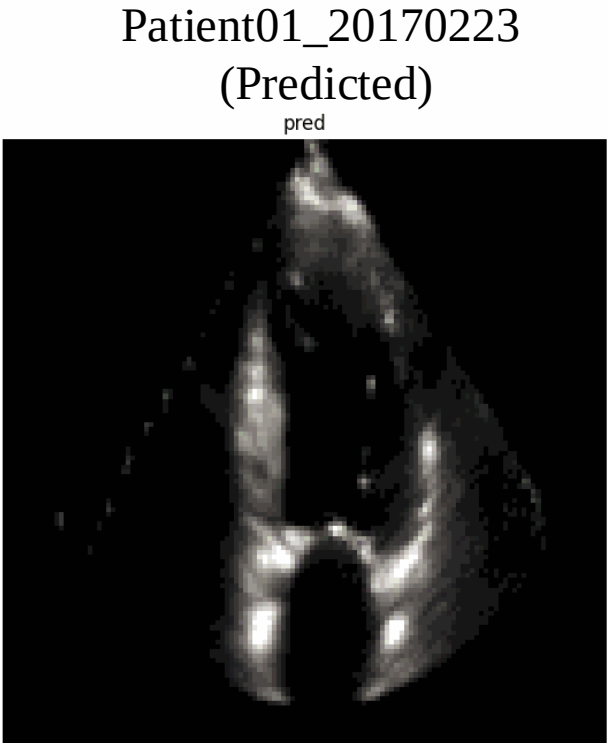
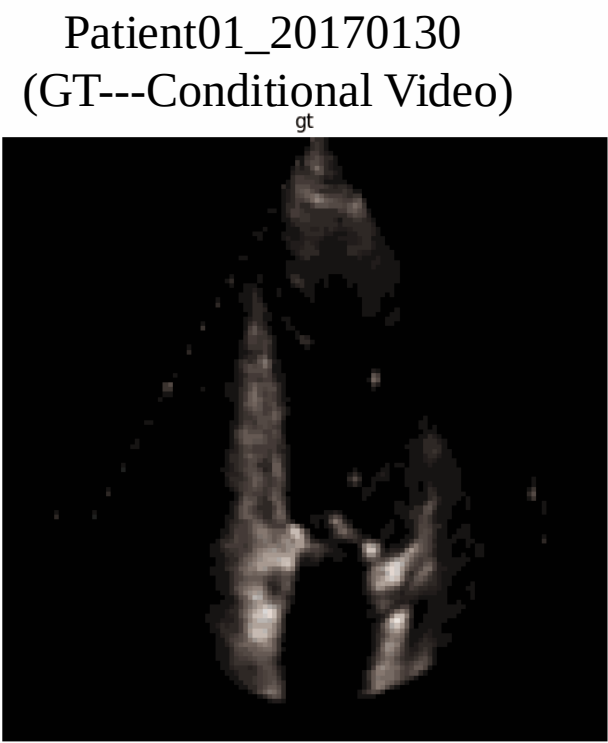
Valvular structural error

Detail blur

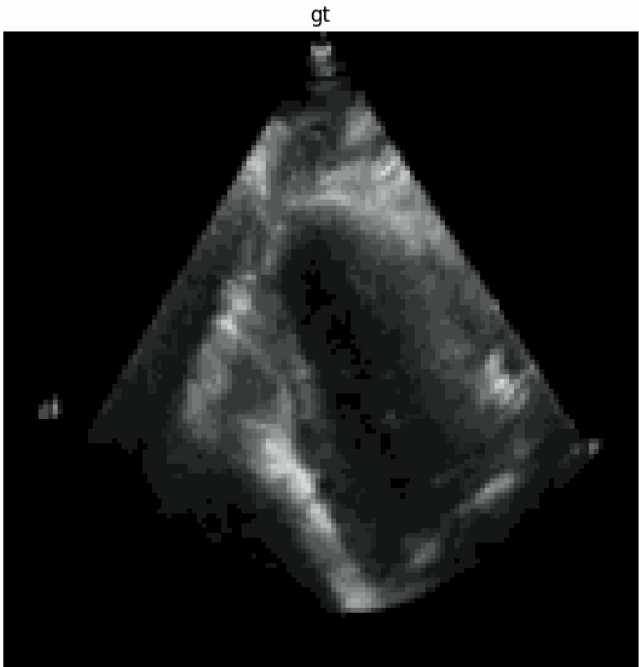


Validation

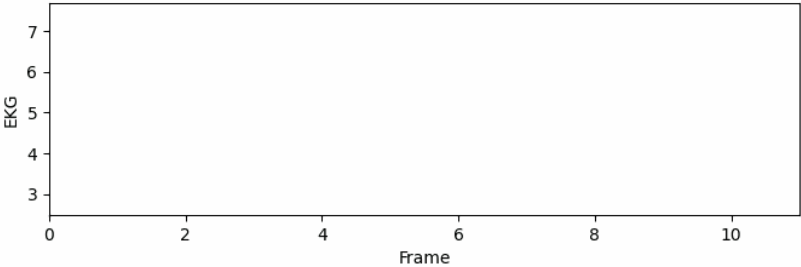
Compare the EF



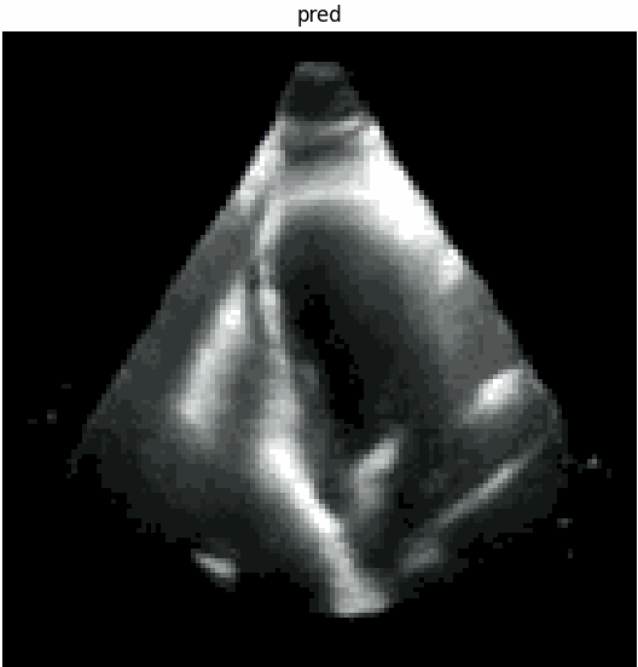
ASM00005 20160209 A2C



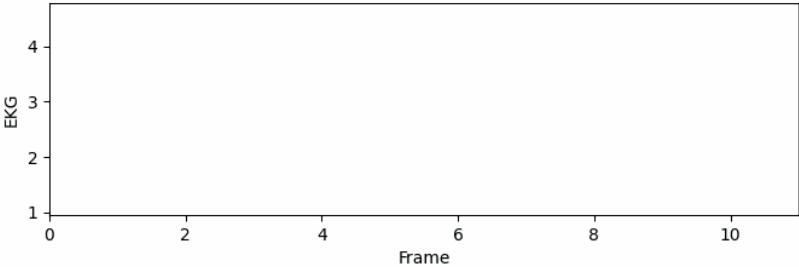
Lead II EKG



ASM00005 20180106 A2C



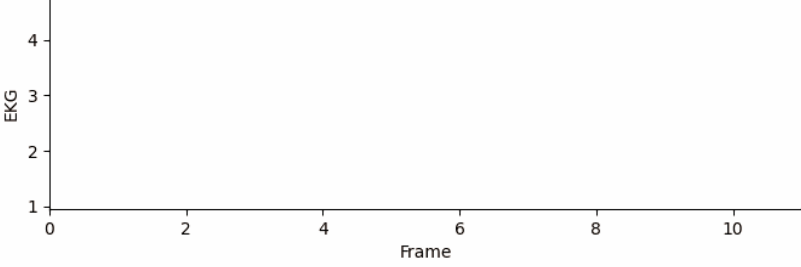
Lead II EKG



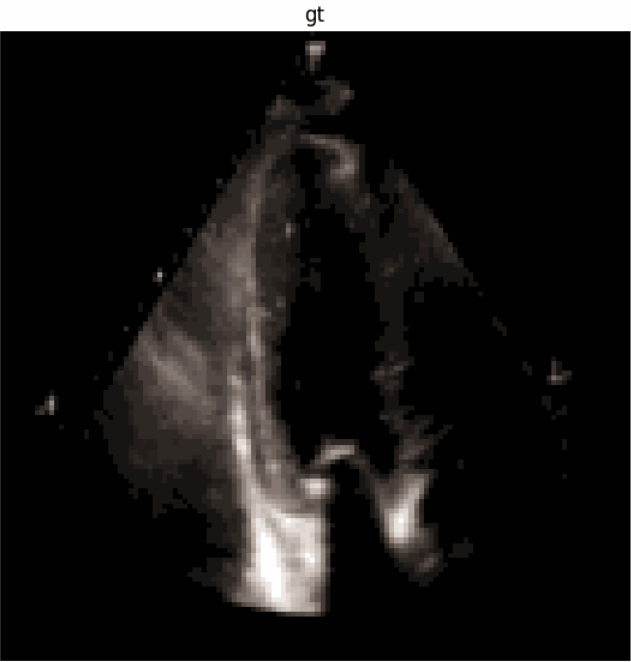
ASM00005 20180106 A2C



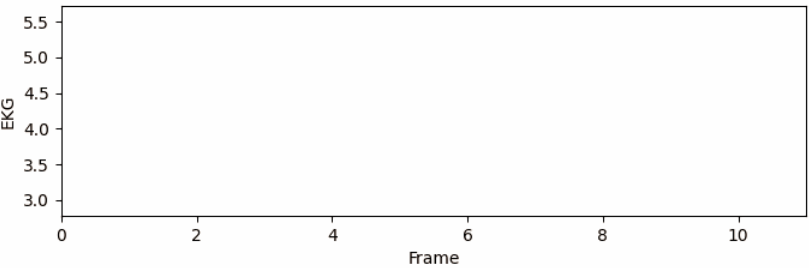
Lead II EKG



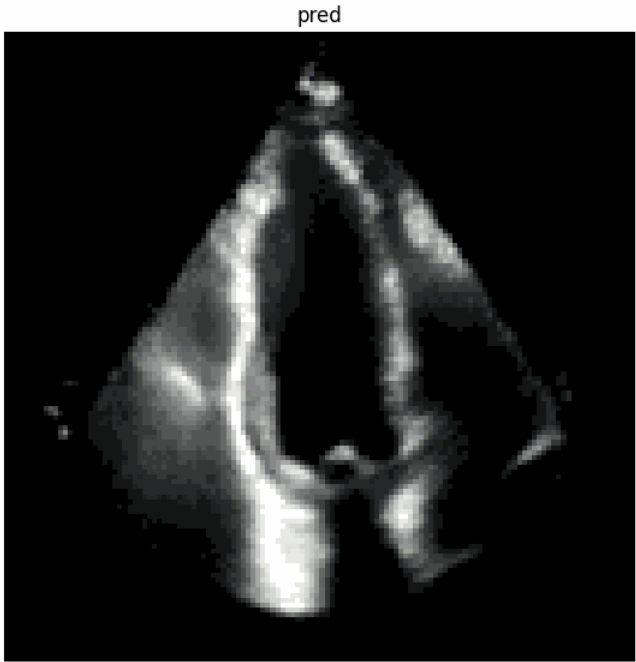
ASM00006 20160619 A3C



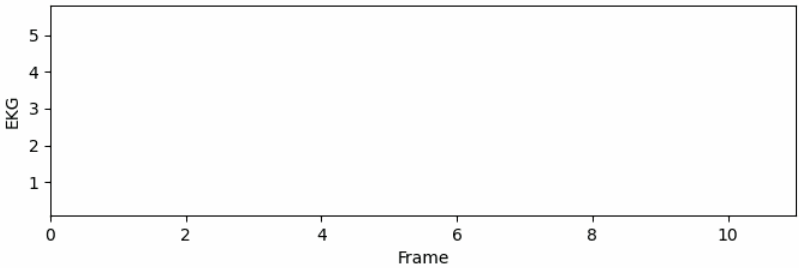
Lead II EKG



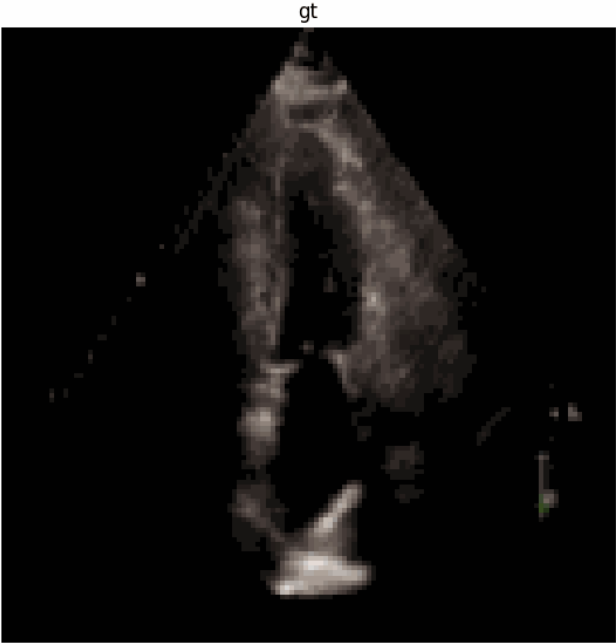
ASM00006 20160701 A3C



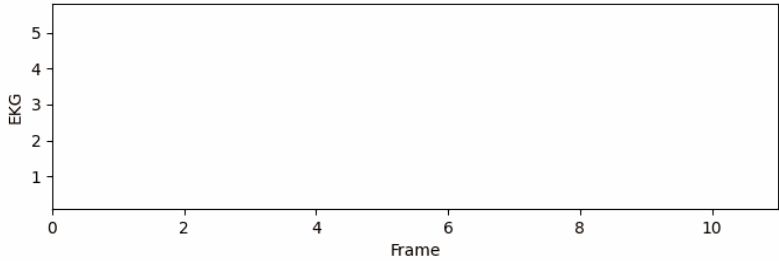
Lead II EKG



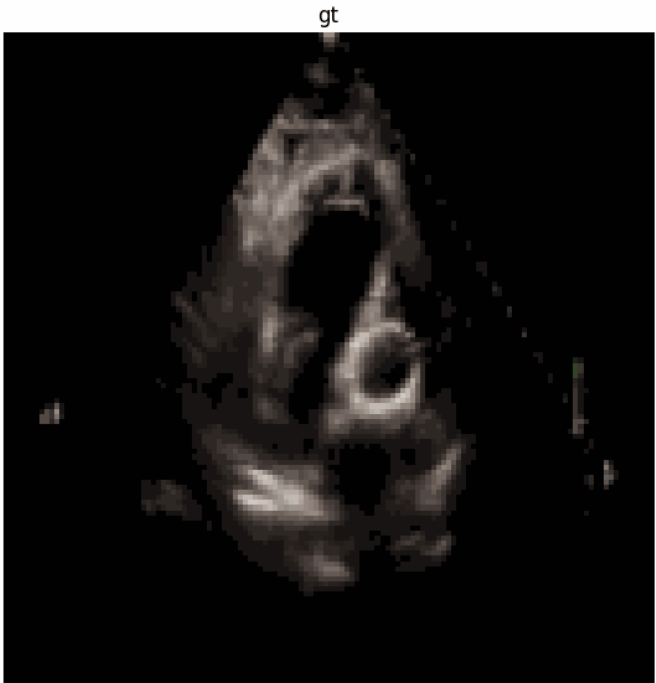
ASM00006 20160701 A3C



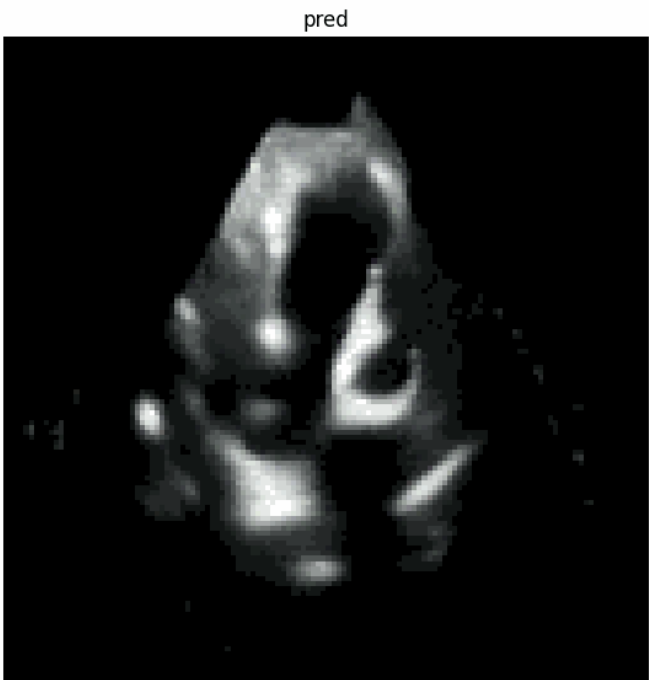
Lead II EKG



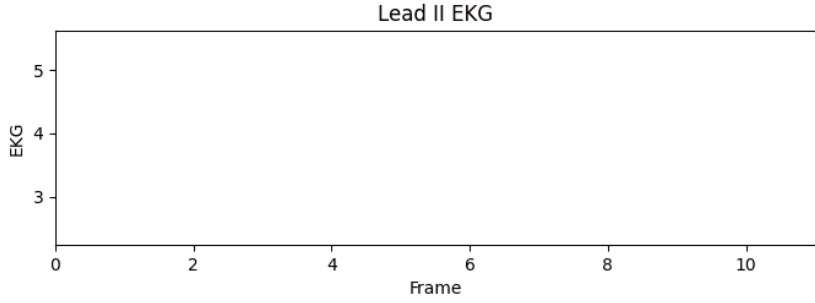
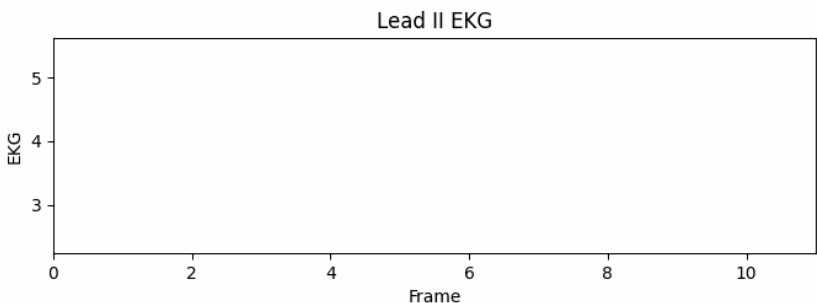
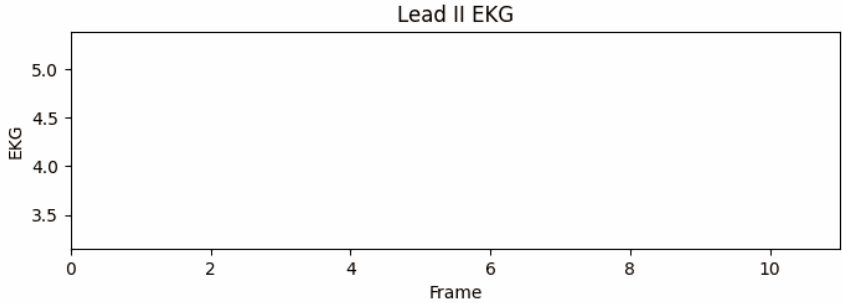
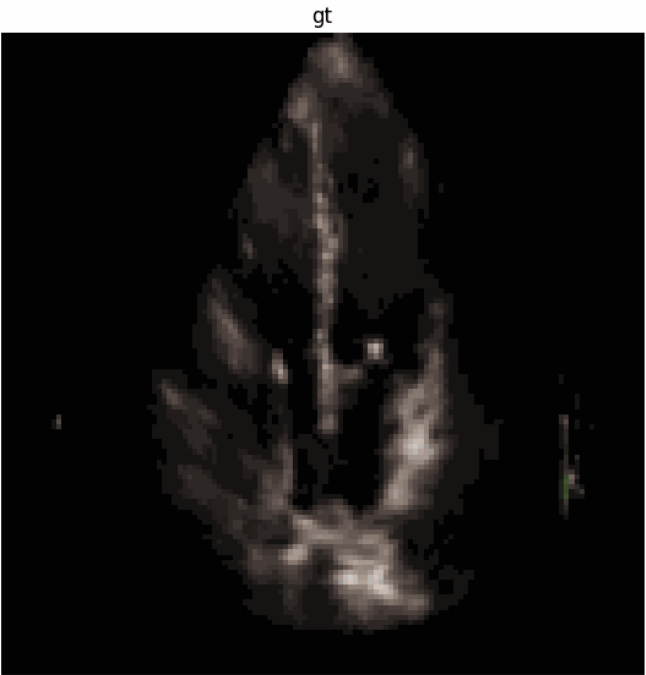
ASM00009 20160406 A4C



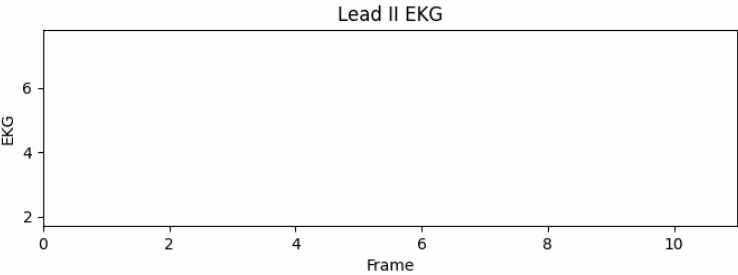
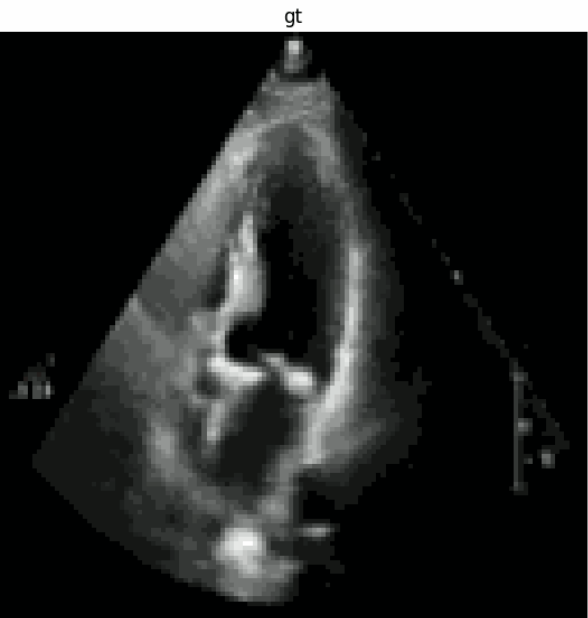
ASM00009 20170504 A4C



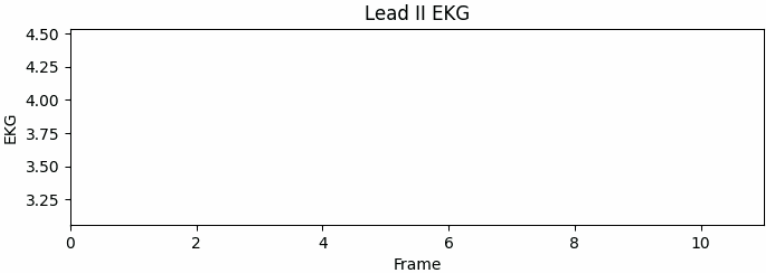
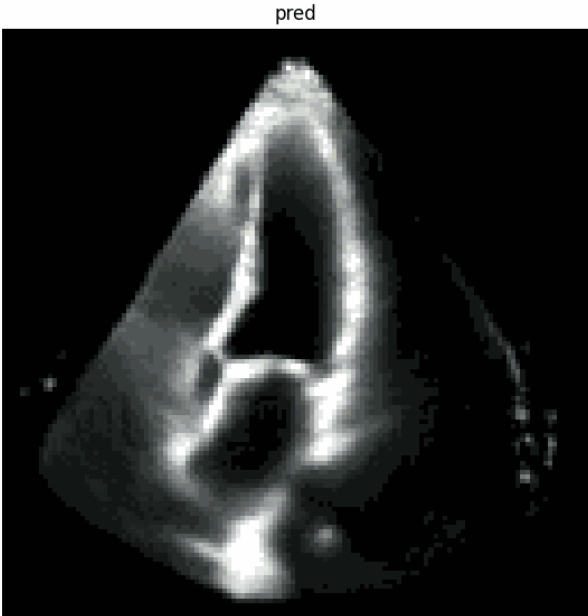
ASM00009 20170504 A4C



ASM00005 20160209 A5C



ASM00005 20180106 A5C



ASM00005 20180106 A5C

