DynamicBUS: Restoring Temporal Dynamics from Static Ultrasound for Improved Breast Cancer Diagnosis

Zhikai Yang[a], Yaofang Liu[b], Tianhao Bai[c], Ander Biguri[d], Haoyuan Chen[e], Yonghao Li[e], Carola-Bibiane Schönlieb[d], Örjan Smedby[a], Raymond Chan[f,*] and Rodrigo Moreno[a]

[a]KTH Royal Institute of Technology, Department of Biomedical Engineering and Health Systems, Stockholm, 14157, Sweden
[b]City University of Hong Kong, Hong Kong, China
[c]Department of Ultrasound, Aviation General Hospital, Beijing, China
[d]University of Cambridge, Cambridge, United Kingdom
[e]ShanghaiTech University, School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, Shanghai, China
[f]Lingnan University, Hong Kong, China

ARTICLE INFO

Keywords:
Video Generation
Breast Ultrasound
Spatio-Temporal Fusion
Diffusion Models
Computer-Aided Diagnosis

ABSTRACT

The clinical practice of archiving static 2D images from dynamic breast ultrasound (BUS) examinations leads to a critical loss of temporal information, vital for accurate lesion diagnosis. This data limitation constrains the performance of conventional computer-aided diagnosis systems. We challenge this limitation by proposing a novel framework that computationally recovers and leverages these lost temporal dynamics to enhance diagnostic accuracy. Our framework first synthesizes a BUS video from a static key frame image using a purpose-built generative model. For diagnostically plausible synthesis, we introduce a key-frame conditioning strategy that ensures the anatomical fidelity of the lesion is preserved while generating useful dynamic cues. Subsequently, the high-fidelity static image and the synthesized video are fed into our designed IV-Net, a dual-branch fusion network that synergistically integrates pristine spatial details with the recovered temporal context for robust classification. Evaluated on key frame BUS datasets, our integrated framework outperforms methods that rely solely on static images (AUC: 92.63%). Moreover, a reader study indicates that the generated videos are indistinguishable to experts and lead to higher diagnostic performance. Overall, our method demonstrates the potential of generative AI to restore lost clinical information, paving the way for more accurate and reliable diagnostic systems in BUS diagnosis. The code and generated dataset are publicly available at https://github.com/minnelab/DynamicBUS

1. Introduction

Breast cancer is the most prevalent and threatening disease in females Sung et al. (2021). Ultrasound imaging is one of the most widely used screening techniques for breast cancer. Radiologists use a transducer across the tissue to observe lesion morphology, echogenicity, and deformation from different points of view. Thus, diagnosing with breast ultrasound (BUS) inherently involves dynamic imaging. This spatio-temporal information is used by clinicians for distinguishing the possibility of malignant and benign lesions. Unfortunately, in clinical practice, only a few representative static 2D frames are usually archived in a patient's health records, while the dynamic video is discarded Evans et al. (2018). This severely limits the capabilities of computer-aided diagnosis (CAD) systems as they only have access to incomplete, static data for training.

While some studies have demonstrated the superior performance of deep learning models on full BUS videos Chen et al. (2021); Sun et al. (2023), these approaches are not viable for widespread clinical use due to the prevalent lack of archived video data. In recent years, different methods have been proposed to generate videos from static images Zhang et al. (2024a); Li et al. (2024). This raises the question of whether these generated breast ultrasound videos can provide additional information for clinical diagnosis.

In this work, we provide a possible answer by proposing a novel, integrated framework that synergistically combines generative video synthesis with a fusion classification network tailored for BUS. The framework consists of two components:

1. **A Temporal Synthesis Module:** We leverage the power of diffusion models to generate a realistic, coherent BUS video from a single static input image. Critically, we introduce a novel key-frame conditioning strategy based on the Frame-Aware Video Diffusion Model (FVDM) Liu et al. (2024b). Unlike conventional methods that condition on the first frame, our approach ensures the central, diagnostically critical frame remains true to the source image, while plausible temporal dynamics are synthesized around it. This approach better resembles the exploration strategy used by clinicians.

2. **A Spatio-Temporal Fusion Classifier:** To effectively utilize this combination of real and synthetic data, we designed IV-Net, a dual-branch network. One branch processes the original, high-fidelity static image to extract precise spatial features, while the other branch analyzes the generated video to capture the synthesized temporal dynamics. To better fuse the image and

*Corresponding author, E-mail address: rhfchan@ln.edu.hk (Raymond Chan) and rodmore@kth.se (Rodrigo Moreno)
**Zhikai Yang and Yaofang Liu contributed equally to this work.
ORCID(s):

video information, we design a bidirectional cross-attention (BCA) mechanism. The information from both modalities is then fused to render a final, more informed classification.

This holistic approach transforms a static image problem into a dynamic video analysis problem in the absence of original video data. To our knowledge, this is the first work to explore and validate the concept of generative temporal recovery for BUS lesion classification.

In summary, our main contributions are:

- We propose a new computer-aided diagnostic paradigm that recovers and leverages lost temporal information from static BUS images, comprising a video generation module and a fusion classification network. To the best of our knowledge, this is the first framework to use synthesized video for BUS classification.
- We introduce a key-frame conditioning strategy for diffusion-based image-to-video (I2V) generation.
- We developed an image-video classification network (IV-Net). This dedicated dual-branch network effectively fuses information from a real static image and a generated video, demonstrating a novel approach to integrating generative data into discriminative tasks.

2. Related Work

2.1. Breast Ultrasound Classification

Within the domain of computer-aided breast ultrasound classification, various methodologies have been extensively explored, including radiomics-based approaches Lee et al. (2018) and deep learning-based methods Shareef et al. (2023); Yang et al. (2024); Gheflati and Rivaz (2022); Byra (2021); Yao et al. (2024); Tang et al. (2025); Iqbal and Sharif (2023). While static BUS image classification has been widely studied, a growing body of evidence highlights the limitations imposed by single-frame analysis. In contrast, methods utilizing BUS videos show improved diagnostic outcomes Chen et al. (2021); Sun et al. (2023); Zhao et al. (2023); Kong et al. (2024); Zhang et al. (2024b). Videos inherently contain more information, as the integration of spatial and temporal features offers greater insight into lesion characteristics like stiffness or motility patterns. However, the practical utility of these video-based models is severely hampered by the clinical data reality, creating a clear need for methods that can bridge this gap.

Deep learning techniques, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and self-attention models, have shown remarkable capabilities in extracting complex patterns from video data Ur Rehman et al. (2023). These models can automatically learn relevant features from sequences of frames, capturing both spatial and temporal characteristics. Our work builds on this potential but obviates the need for pre-existing video datasets by generating the temporal dimension itself.

2.2. Video Diffusion Models

Recent advancements in diffusion models have revolutionized generative modeling, particularly in image synthesis, by employing iterative noise reduction processes to generate high-fidelity samples Song et al. (2020); Ho et al. (2020). While extending this framework to video generation has shown promise, existing models often struggle to capture intricate, long-range temporal dynamics Ho et al. (2022); He et al. (2022); Chen et al. (2023); Wang et al. (2023); Ma et al. (2024); OpenAI (2024); Xing et al. (2023b); Liu et al. (2024a); Hou et al. (2024). For the task of I2V generation, methods like DynamiCrafter Xing et al. (2023a) and I2V-Adapter Guo et al. (2023) have made significant progress, yet often face challenges in preserving high fidelity to the conditioning image while generating realistic motion Zhang et al. (2023); Ni et al. (2024).

In the medical imaging domain, generative models have been explored for applications such as endoscopy, cardiac ultrasound, and fundus video generation Li et al. (2024); Wang et al. (2024); Reynaud et al. (2024, 2023); Yu et al. (2024); Zhang et al. (2024a). However, the unique challenge of generating plausible BUS videos from static images for diagnostic purposes remains an unexplored and critical area of research. Our work is the first to address this specific clinical need, adapting the powerful FVDM Liu et al. (2024b), whose vectorized timestep mechanism is particularly adept at modeling the complex frame interdependencies and supports full preservation of the conditioning image required for this task.

3. Methodology

The proposed framework for BUS classification consists of two main stages: video generation and classification, as shown in Fig. 1. In the first stage, video generation model synthesizes BUS video $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}]$ from the given static image I. Then, the generated video, together with the static image, were used as input to classify the breast tumor, described as follows:

$$\mathbf{X} = \mathcal{G}(I) \tag{1}$$

$$\hat{y} = \mathcal{C}(I, \mathbf{X}) \tag{2}$$

where $\mathcal{G}$, $I \in \mathbb{R}^{w \times h}$, $\mathbf{X} \in \mathbb{R}^{N \times w \times h}$ and $\mathcal{C}$ represent the video generation model, the given image, the generated video and classifier respectively, where N represents the total number of frames in the generated video, w and h denote the height and width. Each image captures a specific lesion or mass region. Given an input static breast ultrasound image I, the diffusion model generates a sequence of frames $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}]$, $X \in \mathbb{R}^{N \times h \times w}$. The $\hat{y}$ is the predicted label (benign or malignant).

In the first stage, we modify FVDM Liu et al. (2024b) to generate a breast ultrasound video from a given static breast image, such that the given image becomes the center frame of the video. In the second stage, we propose an image-video fusion-based classification model (IV-Net) to predict the tumor type.

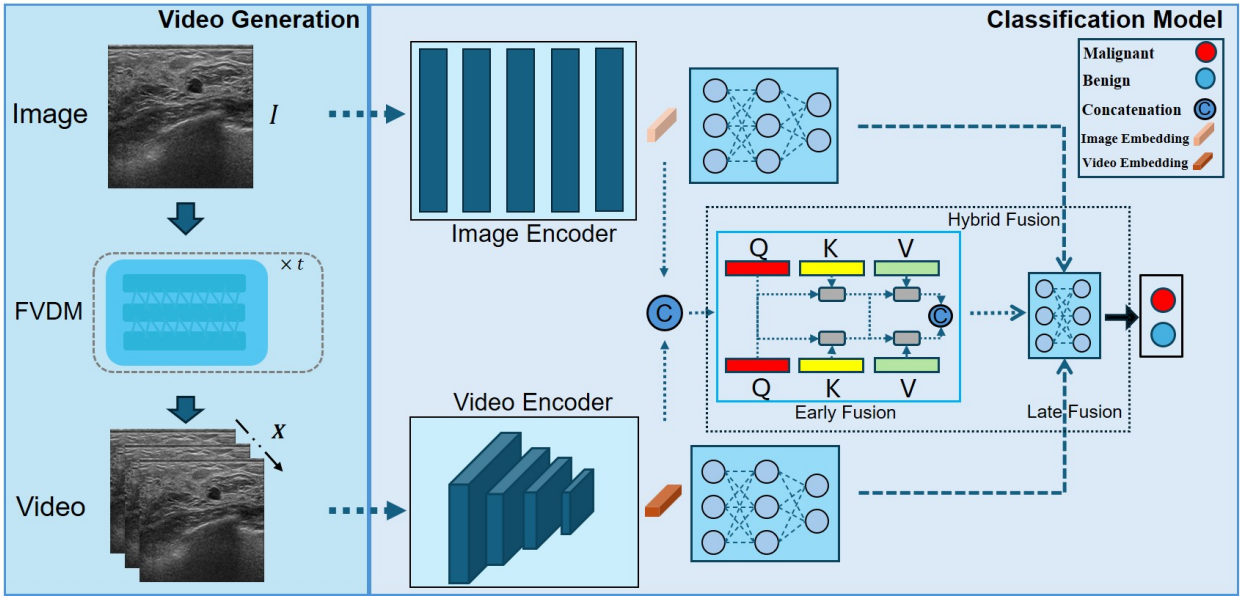

Figure 1: The proposed breast ultrasound classification framework consists of two stages. In the first stage, a diffusion-based model generates a breast ultrasound video from a static image. In the second stage, the generated video and the original static image are used to classify the breast tumor type.

3.1. Key Frame based Breast Ultrasound Video Generation

We present a diffusion based model for synthesizing dynamic BUS videos from static images, addressing the critical need for temporal data in BUS-based cancer classification. The generation model introduces one critical innovation that uses a key frame as a conditional anchor to generate forward and backward frames, diverging from conventional I2V methods Guo et al. (2023); Xing et al. (2023a), which are conditioned on the first frame and generate forward frames. In this way, our model can synthesize ultrasound videos with higher fidelity and better alignment with the conditioning image.

Preliminaries: Diffusion models generate data by reversing a continuous stochastic perturbation process Song et al. (2020). The forward process perturbs data $p_{\text{data}}(\mathbf{x})$ to noise via the Stochastic Differential Equation (SDE):

$$\mathrm{d}\mathbf{x} = \boldsymbol{\mu}(\mathbf{x}, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\mathbf{w}, \tag{3}$$

In this formulation, the drift coefficient $\boldsymbol{\mu}(\mathbf{x}, t)$ governs the deterministic evolution of the data, typically pulling the signal mean towards zero to maintain stability. The diffusion coefficient $\sigma(t)$ controls the rate of stochastic perturbation, scaling the random noise injected by the standard Brownian motion $\mathbf{w}$. Sampling is performed by reversing this dynamics, governed by the reverse-time SDE involving the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$:

$$\mathrm{d}\mathbf{x} = [\boldsymbol{\mu}(\mathbf{x}, t) - \sigma(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\mathrm{d}t + \sigma(t)\mathrm{d}\bar{\mathbf{w}}. \tag{4}$$

Frame-Aware Video Diffusion Model. Standard video diffusion models apply a scalar timestep t uniformly across all frames Xing et al. (2023b), restricting temporal flexibility. To capture complex temporal dynamics, FVDM Liu et al. (2024b) introduces a **vectorized timestep** mechanism $\boldsymbol{\tau}(t) = [\tau^{(1)}(t), \ldots, \tau^{(N)}(t)]^\top \in [0, T]^N$, where each $\tau^{(i)}$ governs the noise level for the i-th frame independently. Consequently, the diffusion dynamics for the i-th frame $\mathbf{x}^{(i)}$ follow the frame-specific SDE:

$$d\mathbf{x}^{(i)} = \boldsymbol{\mu}(\mathbf{x}^{(i)}, \tau^{(i)})dt + \sigma(\tau^{(i)})d\mathbf{w}^{(i)}. \tag{5}$$

By aggregating frames into a video matrix $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}]^\top \in \mathbb{R}^{N \times d}$, we can formulate the integrated forward SDE for the entire video sequence:

$$d\mathbf{X} = \boldsymbol{U}(\mathbf{X}, \boldsymbol{\tau}(t))dt + \boldsymbol{\Sigma}(\boldsymbol{\tau}(t))d\mathbf{W}, \tag{6}$$

where $\boldsymbol{U}$ represents the concatenated drift coefficients, $\boldsymbol{\Sigma}$ is a diagonal matrix of diffusion coefficients determined by $\boldsymbol{\tau}$, and $\mathbf{W}$ denotes the multidimensional Brownian motion.

Key-Frame Anchoring: As mentioned, I2V methods typically generate forward frames from a conditioning image. Instead, we employ the key frame $\mathbf{x}^{(m)}$ as the stable anatomical reference ($m = \lfloor N/2 \rfloor$). Specifically, given a static BUS image I, we initialize:

$$\mathbf{x}_0^{(m)} = I \text{ with } \tau^{(m)}(t) \equiv 0, \tag{7}$$

while applying standard diffusion to other frames:

$$\tau^{(i)}(t) = t, \quad \forall i \neq m. \tag{8}$$

In this way, bidirectional synthesis from the anchor image is enabled, preserving diagnostic fidelity while generating realistic peripheral motion. To be clearer, FVDM supports I2V generation using only the vectorized timestep variable because it allows each frame to evolve independently during the diffusion process. By setting the anchor

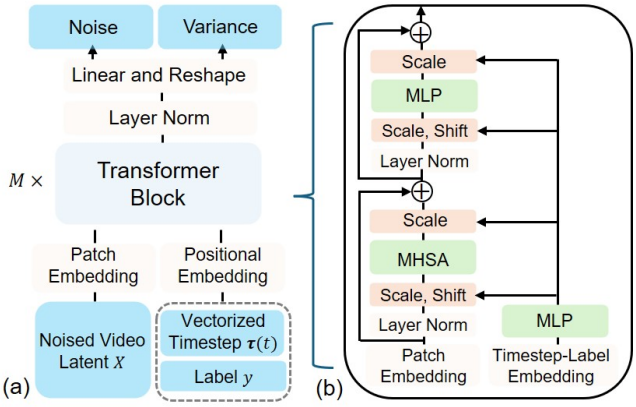

Figure 2: Detailed architecture of the diffusion model. (a) The whole diffusion transformer architecture with components like Patch Embedding, Positional Embedding, Layer Norm, and Transformer Block. (b) Transformer block. MLP and MHSA denote multi-layer perception layer and multi-head self-attention module, respectively.

frame's timestep to zero ($\tau^{(m)}(t) \equiv 0$) and applying regular noise to other frames, the model can treat the input image as a noise-free starting point and generate the other frames from noise. This configuration enables the model to infer motion and temporal dynamics from the static image to create a coherent video sequence without additional training or fine-tuning since the model are trained to handle frames with independent noise levels, which is the prominent advantage of our adopted FVDM paradigm against conventional others.
Model Architecture: Fig. 2 illustrates the full diffusion transformer architecture based on Ma et al. (2024). The difference is replacing the scalar timestep input with a vectorized version. Sinusoidal positional encoding transforms these to embeddings (N, D), where D is the embedding dimension. These embeddings condition the transformer attention and MLP layers via adaLN-Zero Peebles and Xie (2023), enabling independent temporal dynamics per frame for improved temporal fidelity and cross-frame noise prediction. By implementing this enhanced method, we convert static breast ultrasound images into dynamic, multi-frame video sequences.

3.2. Image-Video Classification Network

To better process key frame and generated video information, we propose IV-Net shown in Fig. 1 classification model part. The pair key frame image and generated video together are used as input. The IV-Net consists of three components: an image classification branch, a video classification branch, and a feature fusion module. IV-Net is inspired by the two-branch architecture slow-fast video classification method Feichtenhofer et al. (2019), a model with two streams designed for video understanding tasks. Unlike the slow-fast branches, which capture both slow and fast video sequences, IV-Net's two branches are used to capture the key breast ultrasound image and generate video information. In the image stream, we used ResNet as the backbone to extract the image feature map $f_I \in \mathbb{R}^c$ where c represents the channel dimension. In the video stream, the MViT Li et al. (2022) network model is used for extracting spatial and temporal features $f_X \in \mathbb{R}^c$. After extracting the image and video features, a feature representation fusion module is used to fuse the features.

3.2.1. *Image Video information fusion*

To better incorporate the extracted features, we propose to use the hybrid fusion module to handle the image-video paired data. It combines the early and late fusion module.

Early Fusion In the early fusion strategy, feature-level integration is performed immediately after extracting representations from the image and video branches. We then apply the BCA module to explicitly model interactions between modalities. Fig. 1 shows the early fusion module, which works as follows. First, the features of both branches are projected into a shared latent space $f \in \mathbb{R}^c$ through modality-specific linear layers :

$$q_I = W_I^Q f_I, \quad k_I = W_I^K f_I, \quad v_I = W_I^V f_I,$$

$$q_X = W_X^Q f_X, \quad k_X = W_X^K f_X, \quad v_X = W_X^V f_X.$$

where produces query $W^Q \in \mathbb{R}^{c\times h}$, key $W^K \in \mathbb{R}^{c\times h}$, and value $W^V \in \mathbb{R}^{c\times h}$ representations The image query attends to the video key to obtain video-guided image features, while the video query attends to the image key to obtain image-guided video features. Then, the attended features are computed:

$$\tilde{f}_I = \alpha_I v_X, \quad \tilde{f}_X = \alpha_X v_I,$$

with

$$\alpha_I = \mathrm{Softmax}(q_I k_X^\top), \quad \alpha_X = \mathrm{Softmax}(q_X k_I^\top).$$

Finally, the attended features are concatenated to form the fused representation:

$$f_{\text{early}} = (\tilde{f}_I \| \tilde{f}_X) \in \mathbb{R}^{2c},$$

where $\|$ is a concatenation operation. These features are aggregated and passed through an MLP for classification. By employing bidirectional cross-attention, the model is able to emphasize salient cross-modal interactions and capture complementary dependencies beyond simple concatenation. This attention-based formulation extends the multi-head self-attention mechanism Vaswani et al. (2017) to explicitly model cross-modal interactions at the feature level, allowing the model to emphasize salient information across modalities. This mechanism enables the model to learn complementary information by allowing each modality to selectively attend to the most relevant features from the other, thereby capturing richer cross-modal interactions than simple concatenation or independent processing.

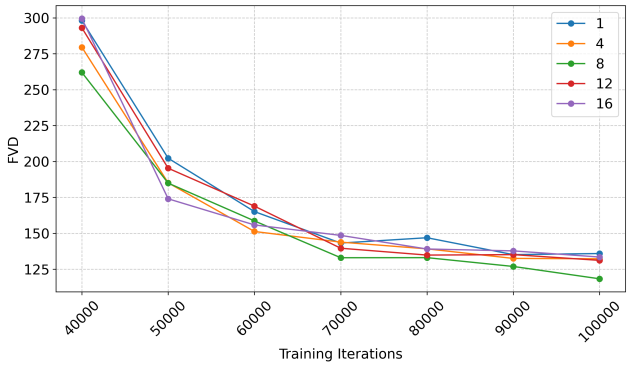

Figure 2: (a) Ablation on Conditioning Position

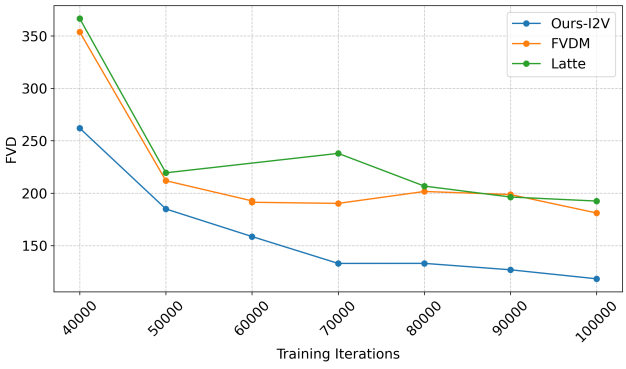

Figure 2: (b) Ablation on Image

Figure 3: (a) Evolution of FVD during training of FVDM w.r.t. different positions for image condition. The green line (conditioning on the 8th frame) achieves the best performance. (b) Evolution of FVD during training of our I2V method (key-anchoring strategy), FVDM Liu et al. (2024b), and Latte Ma et al. (2024).

Late Fusion In contrast, the late fusion strategy defers the integration of information until the decision stage. The image and video features f_I and f_X are individually processed by their respective fully connected layers to obtain independent classification logits:

$$\hat{y}_I = \mathrm{MLP}_I(f_I), \quad \hat{y}_X = \mathrm{MLP}_X(f_X).$$

These logits are then concatenated and fed into a fusion MLP, which acts as a decision-level aggregator:

$$\hat{y} = \mathrm{MLP}_{\mathrm{fusion}}(\hat{y}_I \parallel \hat{y}_X),$$

The fusion MLP consists of three fully connected layers with ReLU activations and dropout for regularization. This structure enables the model to learn a nonlinear mapping from modality-specific decisions to a unified classification output.

The late fusion independent classification logits for image and video features are obtained from their respective fully connected layers. A 3-layer MLP is used to fuse the information at the decision stage, aiming at making decisions based on the results of the two modality branches.

Hybrid Fusion Hybrid fusion integrates both strategies by combining interactions at the feature and decision levels. First, an early fused representation is obtained via BCA:

$$f_{\mathrm{early}} = \mathrm{BCA}(f_I, f_X),$$

This representation is aggregated and projected into a compact vector f_{early}. In parallel, modality-specific features are passed into their respective MLP classifiers to produce logits $\hat{y}_I$ and $\hat{y}_X$. The final hybrid representation is then formed by concatenating the early fused feature and the modality-specific logits:

$$f_{\mathrm{hybrid}} = (f_{\mathrm{early}} \| \hat{y}_I \| \hat{y}_X),$$

which is passed through a fusion MLP to obtain the final output:

$$\hat{y} = \mathrm{MLP}_{\mathrm{fusion}}(f_{\mathrm{hybrid}}).$$

By combining both low-level interactions (early fusion) and high-level decisions (late fusion), this design is expected to provide a more expressive and robust integration mechanism that mitigates the limitations of using either strategy alone.

3.2.2. Loss Function

The binary cross-entropy is used as a loss function: $\mathcal{L}_{\mathrm{BCE}} = -\left[y \log(p) + (1-y)\log(1-p)\right]$, where p denote the prediction probability.

4. Experiment and Results

Our experiments consisted of two parts. The first part evaluated the quality of the generated BUS video sequences by comparing our proposed video-generation framework with existing state-of-the-art approaches. The second part assessed the classification performance of our method by benchmarking it against established BUS classification methods.

4.1. Experiment Setting

The model was implemented in PyTorch and trained on an NVIDIA A100 GPU with 80 GB of memory. We adopted the AdamW optimizer, using a learning rate of $1e^{-4}$ for the diffusion-based generation model and $3e^{-4}$ for the classification model. The generation model was trained from scratch for 100,000 steps with a batch size of 4, while the classification model was trained for 100 epochs with a batch size of 32.

4.1.1. Dataset

In the video generation stage, the Breast Ultrasound Video (BUSV) dataset Lin et al. (2022) was used. It contains a breast lesion ultrasound video dataset with 188 videos, including 113 malignant and 75 benign videos. The 148 cases for training and 40 cases for validation. In total, these videos have 25,272 images. Each video provides a complete scan of the tumor, from its appearance to its largest section and subsequent disappearance. In the classification stage,

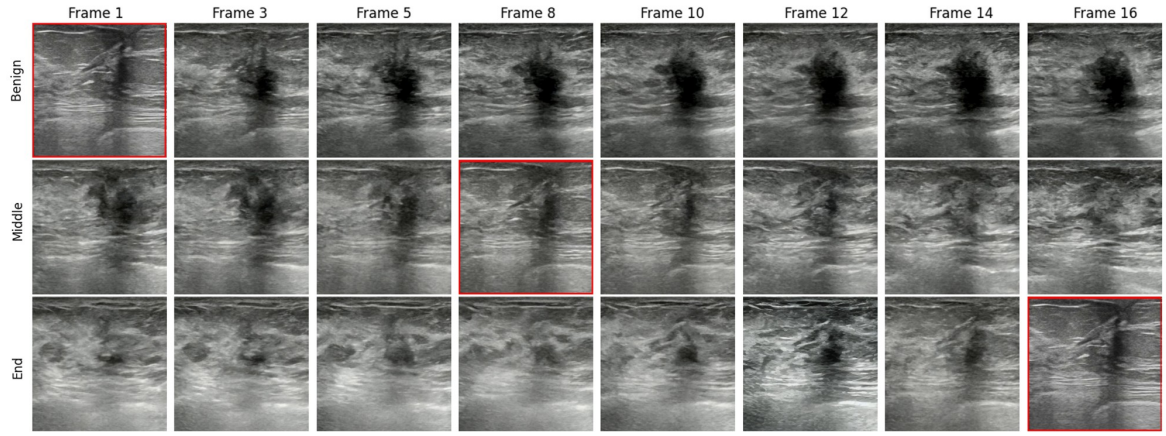

Figure 4: Visualization of the generated video with different condition input. From top to bottom, the generation model takes key images at the beginning, middle, and end positions, which are highlighted with red boxes.

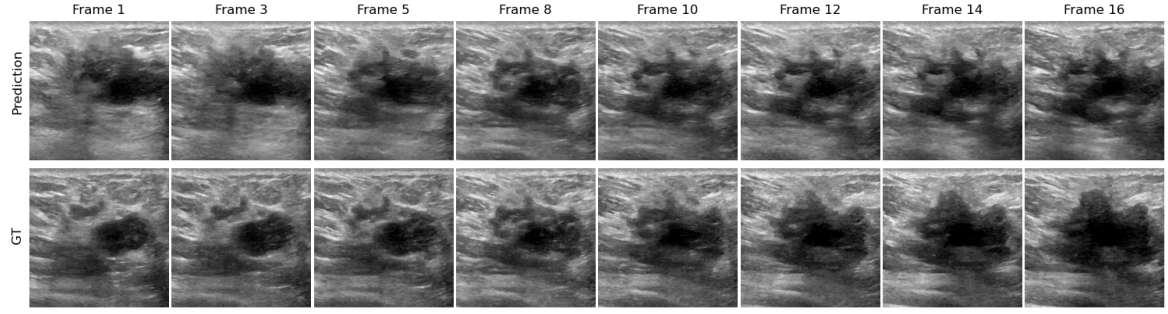

Figure 5: Visualization of our method generated breast ultrasound video sequence compared with the ground truth video sequence.

Table 1
Comparisons with other I2V generation methods.

Method	FVD (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	DISTS (↓)
MCVD Voleti et al. (2022)	639.97	15.79	0.37	0.38	0.24
I2VGen-XL Zhang et al. (2023)	621.92	16.62	0.32	0.48	0.30
SVD-XT Blattmann et al. (2023)	196.92	19.10	0.38	0.36	0.21
LTX-Video HaCohen et al. (2024)	318.56	18.17	0.42	0.46	0.28
CogVideo Hong et al. (2022)	483.86	17.12	0.33	0.45	0.26
Ours	**118.31**	**21.01**	**0.44**	**0.25**	**0.16**

BrEaST Pawłowska et al. (2024) is used, including 252 images in total, 154 benign and 98 malignant images.

Video Generation Dataset Preprocessing: The video generation model is trained with 16 frames of size 256×256 with an interval 3 and without class condition extracted from the BUSV dataset. Given each image in the dataset, we center-crop the image then resize it. After that, we created videos with 16 frames and without class labels.

Classification Dataset Preprocessing: By using the trained video generation model, we generated the video from the static BUS dataset. As the BUS contains heterogeneous noise, we applied a low-pass filter and histogram equalization to enhance image quality. The model was trained on 80% of the dataset, while 20%, consisting of generated videos with known labels, was used for testing.

4.1.2. Metric

For the video generation experiment, the Fréchet Video Distance (FVD) Unterthiner et al. (2019), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018), and Deep Image Structure and Texture Similarity (DISTS) Ding et al. (2020) were used to evaluate the generated video quality.

For the classification experiment, the Precision (PRE), Sensitivity (SEN), the area under the ROC curve (AUC), F1 Score, and Matthews correlation coefficient (MCC) were used for evaluation.

4.2. Video Generation Performance

To determine the optimal configuration of our I2V generation approach, we conducted an ablation study to compare it with baselines and verify the superiority of our proposed approach, as well as a dataset comparison to assess generalization.

4.2.1. Quantitative Comparison

We investigated the effectiveness of our proposed method with ablation studies using the FVD metric (Fig. 3). Fig. 3(a) shows the ablation study for the image conditioned on different frame positions. Specifically, video samples are 16 frames in length. Thus, the conditioning positions range from 1 to 16, representing conditioning from the first frame

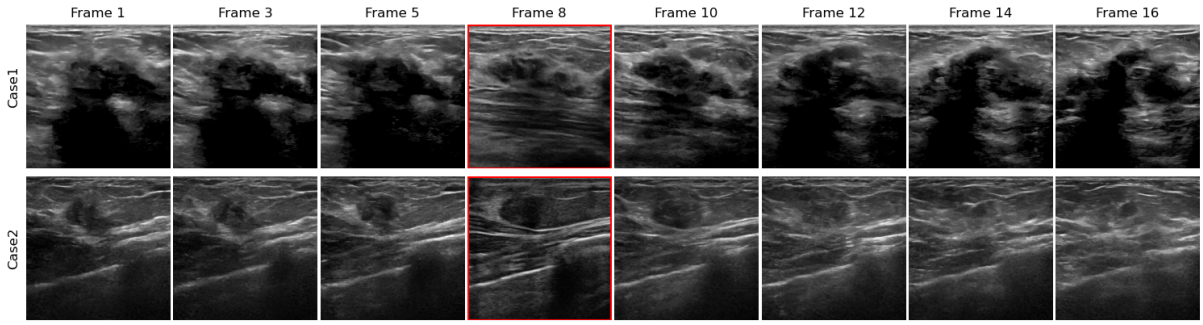

Figure 6: Visualization of videos generated from static images with the proposed method.

to the last. The figure shows that our proposed key-anchoring strategy (conditioning on the middle frame) surpasses all other conditioning positions. Fig. 3(b) compares our I2V generation results with unconditional methods such as original FVDM Liu et al. (2024b) and Latte Ma et al. (2024), which directly generate videos from noise. As shown in Fig. 3(b), the image condition provides more realistic generation results than generating videos from scratch.

We also compared our proposed method with different I2V generation methods, including both zero-shot and specifically trained methods. As shown in Table 1, the results demonstrate that our method outperforms state-of-the-art video generation methods. Specifically, we trained the MCVD model Voleti et al. (2022) on the same dataset from scratch and also list the results for large-scale SOTA models with their pretrained checkpoints. Our method achieved the lowest FVD score of 118.31, indicating higher video quality and better alignment with real video distributions. Additionally, our method attained the highest PSNR (21.01) and SSIM (0.44) scores, signifying superior image fidelity and structural similarity to the ground truth. The model also achieved the lowest LPIPS (0.25) and DISTS (0.16) scores, reflecting enhanced perceptual quality and texture preservation. These results highlight the effectiveness of our proposed approach in bridging the gap between static and dynamic ultrasound imaging.

4.2.2. Qualitative Comparison

We visualized the generated video sequences for comparison, including videos generated with different frame orders, videos from different static image datasets, and a comparison with the ground truth. The image is surrounded by the red line indicating the proposed method's reference input.

Fig. 4 illustrates the generated video with different frame positions as reference inputs. We notice that different conditioning positions result in different generated videos. The proposed key-frame conditioning strategy yields more feasible information that aligns with the original static image, as it can be closer to the original static image from both temporal directions.

Fig. 5 shows the generated BUS video compared to the real video. It demonstrates that the generated video of the tumor area is closer to the ground truth video at frames 10 to 16, recovering most of the original information. Frames from 1 to 5 show more variance, but the overall layout still aligns well.

Fig. 6 visualizes the generated video sequence from the static image dataset. These cases demonstrate that while the tumor is not clearly visible in the static image, the dynamic image reveals a larger range and shows it invading the surrounding area.

4.2.3. Reader Study

To further validate the diagnostic plausibility of the generated BUS videos, we conducted a blinded reader study with an experienced breast ultrasound radiologist (5 years of experience). We performed two studies:

Synthesis Video Quality Evaluation Following a similar procedure applied in a previous study Fu et al. (2023), we designed an experiment to assess whether an expert could distinguish between real BUS videos and those generated by our method, and to evaluate the perceived clinical realism of the synthetic sequences. We sampled a balanced set of 100 paired BUS videos from the BUSV dataset and 100 synthetic videos generated using our method. The distribution of benign and malignant lesions was matched across the two groups. To avoid bias, no real and generated pair from the same lesion was included in the same reader's session.

In the evaluation procedure, each video was presented in a randomized order using a designed experiment platform with standard ultrasound playback controls (start/stop, loop, zoom). For each case, readers were asked to label the video as real or synthetic and rate their confidence on a five-level Likert scale.

The radiologist achieved an average accuracy of 58.2% in distinguishing real from synthetic videos, only slightly above the chance level (50%), with no significant bias toward either class. Confidence was higher when the reader judged videos as real, but both real and synthetic videos received comparable diagnostic quality scores (median Likert rating = 3). These results suggest that our generated BUS videos are largely indistinguishable from real ones in terms of visual plausibility and clinical fidelity.

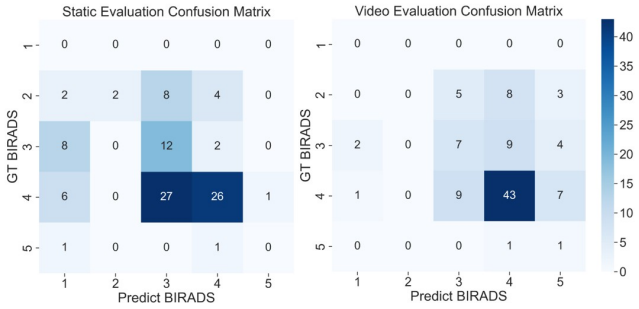

Figure 7: The confusion matrix of the radiologist reader study for predicting BI-RADS scores using static images (left) and adding a generated video (right).

Table 2
Performance of different breast ultrasound classification methods on BUSBRA dataset

Method	Input	PRE	SEN	AUC	F1	MCC (↑)
ViT	Image	76.29	55.46	80.48	62.72	47.14
SwinTrans	Image	78.01	56.93	83.08	65.55	51.31
MViT	Video	77.51	63.50	85.08	67.70	51.24
Ours	Both	**84.86**	**67.41**	**92.63**	**76.15**	**66.18**

Diagnosis Evaluation A diagnostic experiment was conducted to quantify the incremental diagnostic benefit of the generated BUS video compared with a single static image. The experiment was as follows: first, the static breast ultrasound image was presented, and the radiologist was asked to assign a Breast Imaging Reporting and Data System (BI-RADS) score. Then, the generated video of 16 frames was played, and the radiologist was allowed to change the original BI-RADS score. The BrEaST Pawłowska et al. (2024) dataset was used for this experiment.

Fig. 7 shows the radiologist's predicted confusion matrix for static images and breast ultrasound videos. Classes 4A, 4B, and 4C were combined into a single class 4. We found that the overall accuracy improved from 40% with static images to 51% with the image-video pair.

In this experiment, only ultrasound breast images were used, without incorporating clinical information or lesion location data. This limitation resulted in reduced classification performance, as contextual and positional cues are known to contribute to more accurate diagnostic decisions. We acknowledge that the radiologists' BI-RADS performance in our reader study appears relatively low. This can be attributed to the fact that the BI-RADS annotations in the dataset were assigned during the actual clinical examination, where radiologists had access to additional patient information and real-time scanning feedback. In contrast, our reader study restricted radiologists to static images without this supplementary diagnostic context, inherently affecting their performance.

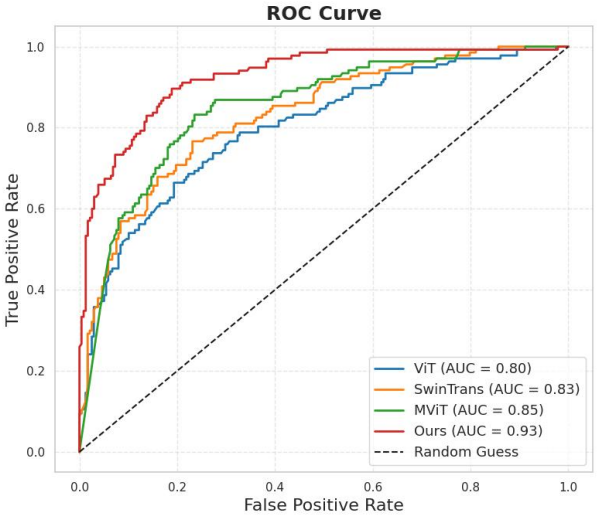

Figure 8: Comparison of the different classification methods ROC curve.

4.3. Comparison of the Classification Performance with Different Methods

To assess the effectiveness of the proposed method, we compared the different breast ultrasound classification models, including image-based and video-based classification models. We compared our proposed IV-BUS method with image-only : Vision Transformer (ViT)Lu et al. (2016), SwinTransLiu et al. (2022), video-only: MViTLi et al. (2022) with pretrained weight. This benchmark model was implemented using the PyTorch library.

As shown in Table 2 and the different methods' ROC curve (Receiver Operating Characteristic), the proposed IV-BUS model performs best. We observed that the proposed method have the best performance. Additionally, the video-based method performs better than the image-based method, which demonstrating that the generated video sequence is helpful for diagnosis. Furthermore, our proposed IV-Net performs better than only image or video-based methods, which shows that it is important to incorporate the original static image for diagnosis.

4.4. Comparison of Classification Performance Under Different Settings

The proposed classification model was evaluated with 4, 8, and 16 frames of the generated video to determine the optimal sequence length for breast ultrasound classification. The 4- and 8-frame sequences were obtained by evenly downsampling the original 16-frame videos. The results in Table 3 indicate that while reducing the number of frames introduces minimal performance degradation, using 16 frames achieves the best overall performance. This suggests that our video generation method effectively captures and generates useful temporal information for classification.

We also investigated different fusion strategies for effectively combining image and video features. As shown in Table 4, we compared early fusion, late fusion, and

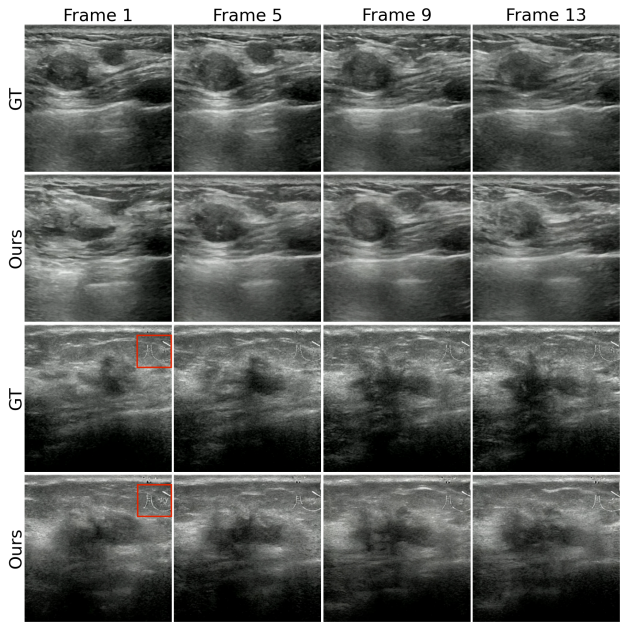

Figure 9: Visualization of limitations in the generated videos. The two rows show a similar comparison for a video with minor color shifts and chart overlays (within the red boxes), where the synthetic video exhibits flickering of chart elements.

Table 3
Classification performance with different numbers of frames for breast ultrasound image-video classification.

Frames	PRE	SEN	AUC	F1	MCC (↑)
4	80.72	66.23	86.69	71.06	57.39
8	81.04	66.67	87.33	71.53	58.09
16	**86.77**	**87.65**	**91.43**	**78.40**	**67.06**

Table 4
Classification performance of different fusion strategies for image-video (Early, Late, Hybrid Fusion)

Method	PRE	SEN	AUC	F1	MCC (↑)
Early	79.48	79.48	83.32	65.81	53.36
Late	85.52	73.68	91.25	78.40	**67.96**
Hybrid	**86.77**	**87.65**	**91.43**	**78.40**	67.06

hybrid fusion mechanisms. The results show that late fusion outperforms early fusion on most metrics, while hybrid fusion achieves the overall best performance, highlighting the importance of flexible feature integration in image-video paired classification.

5. Discussion

We proposed a diffusion-based video generation method that leverages prior BUS video information to assist in diagnosis. This approach offers two key advantages for BUS video generation: (1) preservation of the diagnostic frame at the temporal center and (2) bidirectional anatomical consistency. In this way, static BUS images are transformed into coherent video sequences that capture tissue dynamics critical for malignancy assessment while maintaining diagnostic integrity. To the best of our knowledge, this is the first study to explore the use of generated ultrasound videos for diagnostic purposes, providing a new direction for the application of video generation models in computer-aided diagnosis. We experimented with the generated BUS videos using quantitative and qualitative criteria, and the results show that our proposed method achieves the best video quality.

To assess the value of the generated videos from a clinical perspective, we performed a reader study with an ultrasound radiologist to evaluate both the realism of the videos and whether they could provide additional information for improving diagnosis. We found that the radiologist had low inter-observer variability in the realism reader study. For the diagnosis reader study, the radiologist achieved 51% accuracy with the generated video, compared to 40% with a static image. However, this is still relatively lower than the normal diagnostic procedure. This can be attributed to the fact that the dataset only provides B-mode BUS, whereas BI-RADS diagnosis relies on clinical information, more ultrasound modes (Color Doppler and B-mode), and relative positional information.

We compared our method with several breast ultrasound classification models, and the proposed image–video fusion method consistently achieved the best performance. This demonstrates that accurate video generative models can augment data to provide richer diagnostic cues than static images alone. To identify the optimal integration strategy, we evaluated early, late, and hybrid fusion mechanisms. The results show that hybrid fusion performs best overall, particularly when implemented with a trainable fusion module. We argue that in image–video classification, each modality benefits from independently learning specialized, high-level features before combining them.

Nevertheless, we acknowledge the following limitations. Firstly, the video generation process is based only on a static image without additional information, which may lead to misleading results. In future work, we plan to incorporate more information to aid BUS video generation, such as lesion area masks, text descriptions, and clinical information. Secondly, video generation models may introduce artifacts or misleading information, which could negatively affect diagnostic decisions. Although the motion and structure of the generated videos are well-aligned with real videos, some noticeable visual artifacts can be observed. In Fig. 9, the synthetic video shows minor color shifts and blurring compared to real ultrasound images. Furthermore, if the original video contains overlays such as measurement lines, these elements can appear to flicker in the generated sequence. These issues are common challenges in generative video synthesis and can likely be mitigated by scaling up training data and computational resources, as well as through further algorithmic refinements. In general, despite these minor imperfections, our method shows potential for improving breast ultrasound diagnosis by recovering lost

temporal information. Another issue is that the inference speed of video generation remains a challenge, which is an inherent problem of diffusion models. In future work, we will incorporate anatomical information for video generation and try to accelerate the generation process.

6. Conclusion

In this work, we addressed the critical issue of information loss in breast ultrasound diagnosis, where dynamic video data is routinely discarded in favor of static images. We introduced a diagnostic framework that computationally recovers these lost temporal dynamics to improve classification accuracy. Our approach first synthesizes a realistic BUS video from a static frame using a diffusion model with a keyframe conditioning strategy to ensure anatomical fidelity. Subsequently, a co-designed dual-branch IV-Net synergistically fuses spatial information from the original image with the recovered temporal context from the generated video. Our experiments show that this integrated approach outperforms methods relying on either static images or preexisting videos alone.

Future work will focus on extending our framework to generate longer video sequences, accelerating sampling speed and validating its efficacy across a broader range of datasets. Furthermore, this generative restoration paradigm could be extended beyond BUS to other ultrasound applications where temporal data is routinely lost. This research paves the way for a new class of computer-aided diagnosis systems that can overcome fundamental limitations in clinical data archiving.

Acknowledgments

This work was partially supported by grants from Marie Skłodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01; 101073222), Cancerfonden (22-2389 Pj), HKUST-KTH Global Knowledge Network Awards, HKRGC (grant number CityU11301120, C1013-21GF, CityU11309922), ITF (grant number LU BGR 105824, MHP/054/22), and the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. We thank the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Knut and Alice Wallenberg Foundation for the computational resources at Alvis and Berzelius supercomputers.

References

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al., 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 .

Byra, M., 2021. Breast mass classification with transfer learning based on scaling of deep representations. Biomedical Signal Processing and Control 69, 102828.

Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., Gong, X., 2021. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Transactions on Medical Imaging 40, 2439–2451.

Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al., 2023. Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 .

Ding, K., Ma, K., Wang, S., Simoncelli, E.P., 2020. Image quality assessment: Unifying structure and texture similarity. CoRR abs/2004.07728.

Evans, A., Trimboli, R.M., Athanasiou, A., Balleyguier, C., Baltzer, P.A., Bick, U., Camps Herrero, J., Clauser, P., Colin, C., Cornford, E., et al., 2018. Breast ultrasound: recommendations for information to women and referring physicians by the european society of breast imaging. Insights into imaging 9, 449–461.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6202–6211.

Fu, J., Tzortzakakis, A., Barroso, J., Westman, E., Ferreira, D., Moreno, R., Initiative, A.D.N., 2023. Fast three-dimensional image generation for healthy brain aging using diffeomorphic registration. Technical Report. Wiley Online Library.

Gheflati, B., Rivaz, H., 2022. Vision transformers for classification of breast ultrasound images, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 480–483.

Guo, X., Zheng, M., Hou, L., Gao, Y., Deng, Y., Ma, C., Hu, W., Zha, Z., Huang, H., Wan, P., et al., 2023. I2v-adapter: A general image-to-video adapter for video diffusion models. arXiv preprint arXiv:2312.16693 .

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al., 2024. Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 .

He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q., 2022. Latent video diffusion models for high-fidelity long video generation arXiv:2211.13221.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840–6851.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J., 2022. Video diffusion models. Advances in Neural Information Processing Systems 35, 8633–8646.

Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J., 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 .

Hou, J., Lu, Y., Wang, M., Ouyang, W., Yang, Y., Zou, F., Gu, B., Liu, Z., 2024. A markov chain approach for video-based virtual try-on with denoising diffusion generative adversarial network. Knowledge-Based Systems 300, 112233.

Iqbal, A., Sharif, M., 2023. Bts-st: Swin transformer network for segmentation and classification of multimodality breast cancer images. Knowledge-Based Systems 267, 110393.

Kong, D., Hu, S., Zhao, G., 2024. Mv-stcnet: Breast cancer diagnosis using spatial and temporal dual-attention guided classification network based on multi-view ultrasound videos. Biomedical Signal Processing and Control 87, 105541.

Lee, S.E., Han, K., Kwak, J.Y., Lee, E., Kim, E.K., 2018. Radiomics of us texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma. Scientific reports 8, 1–8.

Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y., 2024. Endora: Video generation models as endoscopy simulators, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 230–240.

Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C., 2022. Mvitv2: Improved multiscale vision transformers for classification and detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4804–4814.

Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L., 2022. A new dataset and a baseline model for breast lesion detection in ultrasound videos, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 614–623.

Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y., 2024a. Evalcrafter: Benchmarking and evaluating

large video generation models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22139–22149.

Liu, Y., Ren, Y., Cun, X., Artola, A., Liu, Y., Zeng, T., Chan, R.H., michel Morel, J., 2024b. Redefining temporal modeling in video diffusion: The vectorized timestep approach. URL: https://arxiv.org/abs/2410.03160, arXiv:2410.03160.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022. Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12009–12019.

Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems 29.

Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y., 2024. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 .

Ni, H., Egger, B., Lohit, S., Cherian, A., Wang, Y., Koike-Akino, T., Huang, S.X., Marks, T.K., 2024. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9015–9025.

OpenAI, 2024. Sora: Creating video from text. https://openai.com/sora .

Pawłowska, A., Ćwierz-Pieńkowska, A., Domalik, A., Jaguś, D., Kasprzak, P., Matkowski, R., Fura, Ł., Nowicki, A., Żołek, N., 2024. Curated benchmark dataset for ultrasound based breast lesion analysis. Scientific Data 11, 148.

Peebles, W., Xie, S., 2023. Scalable diffusion models with transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4195–4205.

Reynaud, H., Meng, Q., Dombrowski, M., Ghosh, A., Day, T., Gomez, A., Leeson, P., Kainz, B., 2024. Echonet-synthetic: Privacy-preserving video generation for safe medical data sharing, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 285–295.

Reynaud, H., Qiao, M., Dombrowski, M., Day, T., Razavi, R., Gomez, A., Leeson, P., Kainz, B., 2023. Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 142–152.

Shareef, B.M., Xian, M., Sun, S., Vakanski, A., Ding, J., Ning, C., Cheng, H.D., 2023. A benchmark for breast ultrasound image classification. Available at SSRN 4339660 .

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B., 2020. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 .

Sun, A., Zhang, Z., Lei, M., Dai, Y., Wang, D., Wang, L., 2023. Boosting breast ultrasound video classification by the guidance of keyframe feature centers, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 441–451.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 71, 209–249. URL: https://doi.org/10.3322/caac.21660, doi:10.3322/caac.21660.

Tang, L., Zhou, Q., Zhang, Y., Hou, N., Zhu, G., 2025. Lra-unet: Large receptive-field attention-enhanced u-net for breast lesion segmentation in ultrasound images. Knowledge-Based Systems , 114890.

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S., 2019. Fvd: A new metric for video generation .

Ur Rehman, A., Belhaouari, S.B., Kabir, M.A., Khan, A., 2023. On the use of deep learning for video classification. Applied Sciences 13, 2007.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Voleti, V., Jolicoeur-Martineau, A., Pal, C., 2022. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in neural information processing systems 35, 23371–23385.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S., 2023. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 .

Wang, Z., Zhang, L., Wang, L., Zhu, M., Zhang, Z., 2024. Optical flow representation alignment mamba diffusion model for medical video generation. arXiv preprint arXiv:2411.01647 .

Xing, J., Xia, M., Zhang, Y., Chen, H., Wang, X., Wong, T.T., Shan, Y., 2023a. Dynamicrafter: Animating open-domain images with video diffusion priors. arXiv preprint arXiv:2310.12190 .

Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G., 2023b. A survey on video diffusion models. arXiv preprint arXiv:2310.10647 .

Yang, Z., Fan, T., Smedby, Ö., Moreno, R., 2024. 3d breast ultrasound image classification using 2.5 d deep learning, in: 17th International Workshop on Breast Imaging (IWBI 2024), SPIE. pp. 443–449.

Yao, Y., Duan, X., Qu, A., Chen, M., Chen, J., Chen, L., 2024. Dfcg: A dual-frequency cascade graph model for semi-supervised ultrasound image segmentation with diffusion model. Knowledge-Based Systems 300, 112261.

Yu, J., Chen, R., Zhou, Y., Chen, Y., Duan, Y., Huang, Y., Zhou, H., Tan, T., Yang, X., Ni, D., 2024. Explainable and controllable motion curve guided cardiac ultrasound video generation, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 232–241.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR.

Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J., 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 .

Zhang, W., Huang, S., Yang, J., Chen, R., Ge, Z., Zheng, Y., Shi, D., He, M., 2024a. Fundus2video: Cross-modal angiography video generation from static fundus photography with clinical knowledge guidance, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 689–699.

Zhang, Y., Kong, D., Li, J., Yang, T., Yao, F., Yang, G., 2024b. Using segment-level attention to guide breast ultrasound video classification, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Zhao, G., Kong, D., Xu, X., Hu, S., Li, Z., Tian, J., 2023. Deep learning-based classification of breast lesions using dynamic ultrasound video. European Journal of Radiology 165, 110885.