OrgaCast: A Trustworthy Spatiotemporal Diffusion Model for Fluorescence Organoid Forecasting

Dawei Gao*, Angello Huerta Gomez*, Mingchen Li, Marcel El-Mokahal, Huaxiao Yang†, Yunhe Feng†

University of North Texas
Denton, TX, USA
{Dawei.Gao, Angello.HuertaGomez, Mingchen.Li, Marcelel-Mokahal, Huaxiao.Yang, Yunhe.Feng}@unt.edu

Abstract

Accurately forecasting the spatiotemporal dynamics of biological systems, such as human pluripotent stem cell (hPSC)-derived cardiac organoids, from microscopy time-series is a critical challenge in biomedicine with profound implications for drug discovery. Existing generative models often fail to capture the intricate dynamics of organoid development, struggling with their irregular morphology, indistinct boundaries, and complex spatiotemporal patterns. To overcome these limitations, we introduce *OrgaCast*, a novel multimodal conditional diffusion model for high-fidelity organoid forecasting. *OrgaCast* uniquely conditions the generative process on three synergistic modalities: (i) historical image sequences, captured by a dedicated spatiotemporal control module; (ii) structured numerical metadata defining experimental conditions; and (iii) descriptive text captions summarizing the biological context. This comprehensive conditioning enables the generation of forecasts with high visual accuracy and biological plausibility. Furthermore, to enhance the model's utility in critical research settings, we introduce a post-hoc uncertainty quantification method that produces intuitive confidence maps, bolstering the interpretability and trustworthiness of predictions. Extensive experiments on a challenging cardiac organoid dataset demonstrate that *OrgaCast* outperforms baselines in metrics such as SSIM, PSNR, and LPIPS. Our framework presents a robust solution for biological forecasting, promising to accelerate research discovery while minimizing experimental costs and manual effort.

1 Introduction

Accurate forecasting of the developmental dynamics in human pluripotent stem cell (hPSC)-derived cardiac organoids, captured through time-lapse fluorescence microscopy, is crucial for advancing cardiovascular biomedical research and accelerating drug discovery (Bian et al. 2021; Van Valen et al. 2016; Esteva et al. 2019). These cardiac organoids represent New Approach Methodologies (NAMs) for modeling human heart development and diseases *in vitro* (Beilmann et al. 2025; Kandula et al. 2025; Abilez et al. 2025). Fluorescence microscopy image sequences offer valuable insights into key biological phenomena, including cellular proliferation, differentiation, growth, and tissue self-organization. Consequently, predictive models capable of reliably forecasting future organoid states, such as structural and cellular composition, hold great potential for reducing the dependency on costly and labor-intensive wet-lab procedures, such as fluorescence immunostaining and microscopic imaging (Tiwari et al. 2021; Sermesant et al. 2021).

Recent advancements in generative modeling, particularly diffusion-based video synthesis methods employing three-dimensional convolutions and temporal attention, have shown impressive performance in natural video generation tasks (Ho et al. 2022). However, these models struggle when applied directly to fluorescence microscopy data due to the unique domain-specific challenges (Özbey et al. 2023). In contrast to natural scenes, fluorescence microscopy images often feature multiple spectral channels that individually highlight distinct biological structures. Moreover, organoids exhibit blurred boundaries, overlapping cellular structures, and low signal-to-noise ratios within typically dark and uniform backgrounds (Lehtinen et al. 2018). These inherent characteristics significantly reduce the effectiveness of conventional diffusion-based methods, which are optimized primarily for clear and well-defined visual contexts.

Furthermore, visual information alone frequently proves inadequate for reliable organoid forecasting (de Medeiros et al. 2022). Similar visual morphologies can emerge under distinct experimental conditions, creating ambiguities that purely visual models are ill-equipped to resolve (Du et al. 2023). To overcome this limitation, incorporating additional contextual information becomes essential (Zhang, Rao, and Agrawala 2023). Such information encompasses structured numerical metadata, such as cell line identifiers, treatment specifics, and time points, as well as descriptive textual summaries of biological phenomena (Hammer et al. 2021; Papargyriou et al. 2024). Drawing inspiration from expert interpretation practices in biological microscopy, we posit that a successful forecasting model must integrate multimodal data to enhance both predictive accuracy and interpretability (Wang et al. 2024).

In response to these challenges, we introduce *OrgaCast*, a novel multimodal conditional diffusion framework explicitly designed to accurately model dynamic biological pro-

*These authors contributed equally.
†Corresponding author.

cesses with enhanced fidelity and trustworthiness. *OrgaCast* operates in a latent diffusion space for computational efficiency and introduces two core architectural innovations. First, it employs a dedicated 3D spatiotemporal control module to explicitly model the complex temporal dependencies from historical image sequences. Second, it integrates this visual context with conditioning from structured metadata and descriptive text prompts, allowing it to resolve ambiguity and generate predictions that are both visually accurate and biologically plausible.

Beyond predictive accuracy, a crucial requirement for models in high-stakes domains is trustworthiness. To ensure generated predictions are visually realistic and scientifically reliable, we propose a hierarchical multi-scale loss function that optimally balances pixel-level reconstruction accuracy, perceptual quality, and structural consistency. Additionally, we introduce a post-hoc uncertainty quantification module designed to generate confidence maps, enabling users to critically assess prediction reliability and highlight specific areas requiring further expert examination. In summary, our primary contributions are as follows:

- We propose *OrgaCast*, a novel multimodal generative model specifically tailored for accurate forecasting of cardiac organoid development. By integrating historical imaging sequences, structured metadata, and textual descriptions, *OrgaCast* provides biologically plausible predictions in fluorescence microscopy.
- We introduce a hierarchical multi-scale loss function designed explicitly to enhance visual accuracy, perceptual realism, and structural integrity in predicted images.
- We develop an uncertainty-aware inference module capable of generating confidence maps, significantly enhancing prediction interpretability and facilitating targeted expert analyses.
- Extensive experimental evaluations on cardiac organoid datasets demonstrate *OrgaCast*'s superior performance over existing methods, as quantified by metrics such as SSIM, PSNR, and LPIPS.

2 Related Work

In this section, we outline generative spatiotemporal forecasting, review multimodal conditioning strategies, and survey biomedical generative models with emphasis on organoid-scale fluorescence microscopy.

2.1 Generative Spatiotemporal Forecasting

Forecasting future observations in spatiotemporal data is a fundamental problem in scientific modeling, with applications ranging from video prediction to biomedical imaging. Traditional methods like recurrent neural networks (RNNs) model sequences step-by-step but struggle with long-range patterns (Wang et al. 2017). More recently, transformer-based architectures have gained popularity due to their ability to capture global temporal dependencies. For example, TimeSformer applies divided space-time attention for video understanding (Bertasius, Wang, and Torresani 2021), while Video Swin Transformer employs hierarchical spatiotemporal attention to model local and global dynamics (Liu et al. 2022). Diffusion models represent a newer class of generative forecasting methods that learn complex spatiotemporal distributions in a stable, iterative manner (Ho et al. 2022; Yang et al. 2023). For instance, Video Diffusion Models (VDM) generate temporally coherent sequences using 3D convolutions and time-aware attention (Ho et al. 2022). However, these models are predominantly trained on natural video datasets with regular frame intervals, sharp object boundaries, and rich textures. In scientific domains such as time-lapse fluorescence microscopy, the data characteristics differ substantially from those in natural video models. Biomedical image sequences often suffer from irregular sampling, blurry boundaries and overlapping structures (Yilmaz, Eschweiler, and Stegmaier 2024). Therefore, forecasting in biomedical domains demands models that are specifically designed to accommodate their unique structural and temporal properties (He et al. 2025).

2.2 Multimodal Generative Conditioning

Modern generative models increasingly rely on multimodal inputs to improve controllability and output quality. This trend is especially visible in text-to-image models such as Stable Diffusion(Saharia et al. 2022), where natural language prompts are used to guide image generation. ControlNet (Zhang, Rao, and Agrawala 2023) extends this paradigm by injecting structural priors such as edge maps and segmentation masks into the diffusion process, enabling fine-grained control over spatial layout. In scientific applications where visual information may be ambiguous or incomplete, multimodal conditioning becomes especially important (Mou et al. 2024). Structured metadata, which contains essential experimental details such as treatment conditions and imaging time points, can help disambiguate visually similar patterns and enhance predictive accuracy and interpretability (Hammer et al. 2021). Recent models like SpotDiff (Chen et al. 2025) demonstrate that combining spatial transcriptomics with textual prompts improves downstream imputation. In our context, diffusion-based forecasting can similarly benefit from integrating multiple modalities.

2.3 Generative Biomedical Forecasting

Biomedical image sequences, such as time-lapse fluorescence microscopy data, present unique challenges for generative modeling due to irregular sampling, blurry boundaries, and overlapping structures (Yi, Walia, and Babyn 2019). Denoising Diffusion Probabilistic Models (DDPMs) (Dhariwal and Nichol 2021) have shown state-of-the-art performance in image synthesis by modeling complex data distributions through iterative denoising. These advances have inspired applications across both image-based and genomic domains. For instance, SpotDiff (Chen et al. 2025) extends diffusion to spatial transcriptomics by conditioning generation on both gene expression data and learned textual prompts, demonstrating the potential of multimodal guidance in scientific contexts. (Wolleb et al. 2022) proposed a diffusion-based framework for medical anomaly detection by learning distributions of healthy anatomical structures and identifying deviations in medical scans. (Pinaya et al. 2022) applied latent diffusion models to generate high-fidelity syn-

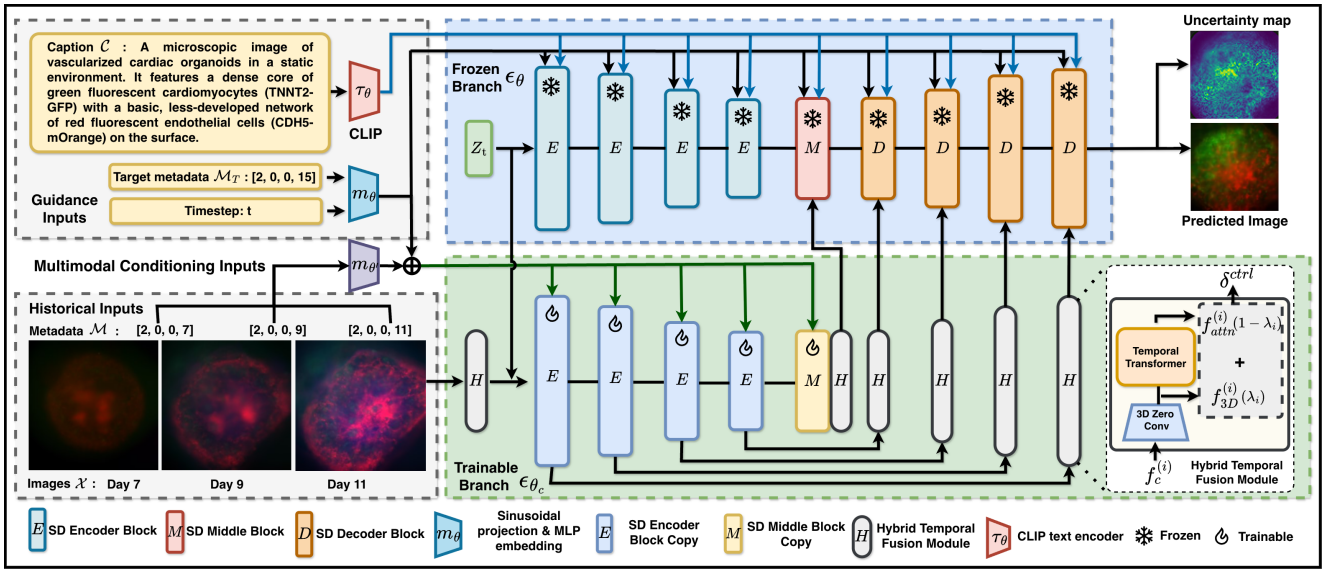


Figure 1: Overview of the *OrgaCast* architecture. The model conditions its generation process on three sources: (1) historical images $\mathcal{X}$, which are processed by a trainable spatiotemporal control branch (ϵ_{θ_c}) to generate an additive control signal δ^{ctrl}; (2) structured metadata $\mathcal{M}$, which is embedded into a vector c_{meta} to condition the diffusion timestep t; and (3) text captions $\mathcal{C}$, which are encoded via a CLIP encoder into c_{text} to guide generation through cross-attention. These signals are fused within a frozen U-Net backbone (ϵ_θ) to generate visually realistic and biologically meaningful predictions.

thetic brain images, primarily for data augmentation and neuroimaging analysis. However, to the best of our knowledge, no existing diffusion-based framework jointly models spatial structures, temporal progression, and contextual metadata for predictive forecasting of biomedical image sequences. We address this research gap by proposing *OrgaCast*, a multimodal diffusion model that integrates historical image sequences, structured experimental metadata, and natural language descriptions to generate biologically plausible predictions of future cardiac organoid states.

3 Methodology

Our goal is to develop a trustworthy, biologically informed generative framework for forecasting cardiac organoid states, formulated as a conditional diffusion process with a multimodal architecture and hierarchical objectives.

3.1 Problem Formulation

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$ denote a time-ordered sequence of fluorescence microscopy images of cardiac organoids, where each $x_t \in \mathbb{R}^{H \times W \times N}$ represents a multi-channel image captured at time point t. Here, H and W correspond to the image height and width, respectively, and N denotes the number of fluorescence channels. In this study, we consider three channels ($N = 3$), with each channel highlighting a specific cell type: the green channel (TNNT2-GFP) labels cardiomyocytes, the red channel (CDH5-mOrange) labels endothelial cells, and the blue channel (TAGLN-CFP) labels smooth muscle cells. The sequence length T reflects the number of historical observations available for a given cardiac organoid sample.

Given the observed image sequence $\mathcal{X}$, structured experimental metadata $\mathcal{M}$ (e.g., cell line, treatment conditions, imaging time points), and a natural language description $\mathcal{C}$ (capturing high-level semantic context such as image type, developmental stage, and morphological features), our objective is to predict a future fluorescence image x_{T+k} at a subsequent time point $T + k$. Formally, the task is to learn a conditional generative mapping $(\mathcal{X}, \mathcal{M}, \mathcal{C}) \rightarrow \hat{x}_{T+k}$, where $\hat{x}_{T+k}$ denotes the forecasted image. This problem constitutes a spatiotemporal generative forecasting task, which requires modeling both the spatial organization within individual fluorescence images and the temporal dynamics across the image sequence. We approach this task through a conditional diffusion process, enabling our model to generate biologically plausible and context-aware predictions of future organoid development.

3.2 *OrgaCast* Framework Overview

We introduce *OrgaCast*, a novel multimodal conditional diffusion framework purpose-built for predictive modeling of organoid fluorescence microscopy images. Illustrated in Figure 1, *OrgaCast* innovatively conditions the generative process on three synergistic modalities: (1) $\mathcal{X}$: a temporal sequence of historical fluorescence images capturing organoid development, (2) $\mathcal{M}$: structured numerical metadata delineating experimental parameters and developmental timepoints, and (3) $\mathcal{C}$: descriptive natural language captions summarizing relevant biological context.

At its core, *OrgaCast* utilizes a hybrid architecture that seamlessly fuses spatiotemporal visual sequences with structured metadata and semantic context. Given these multimodal inputs, *OrgaCast* outputs both the predicted future

fluorescence image $\hat{x}_{T+k}$ and a corresponding uncertainty map U. This integrated design empowers the model to synthesize future organoid images with enhanced visual fidelity, biological realism, and strict adherence to experimental conditions. By combining complementary sources of information, *OrgaCast* offers trustworthy and interpretable forecasts that accurately capture the intricate processes of cellular differentiation and tissue self-organization fundamental to organoid development.

3.3 Model Architecture and Conditioning

Our framework, *OrgaCast*, is built upon a latent diffusion model (LDM), which we augment with a novel multimodal conditioning architecture designed for spatiotemporal forecasting. We detail its three core components: the LDM backbone, the spatiotemporal control module, and the side-channel conditioning mechanism.

Latent Diffusion Backbone Following standard LDM practice, we first encode an image x into a lower-dimensional latent representation z_0. The forward diffusion process gradually adds Gaussian noise to this latent variable over T_{diff} discrete timesteps according to a predefined variance schedule $\bar{\alpha}_t$. At any timestep $t \in \{1, \dots, T_{\text{diff}}\}$, the noisy latent z_t is obtained in a closed form:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$$

The reverse process is learned by a denoising U-Net, denoted $\boldsymbol{\epsilon}_\theta(z_t, t, \mathbf{c})$, which is trained to predict the noise $\boldsymbol{\epsilon}$ from the noisy latent z_t, the timestep t, and a comprehensive conditioning vector $\mathbf{c}$.

Spatiotemporal Control via 3D-Conditioning Module To inject historical context from the image sequence $\mathcal{X}$, we introduce a powerful spatiotemporal control module. Our design extends the dual-branch principle of ControlNet (Zhang, Rao, and Agrawala 2023) to the temporal domain, an approach inspired by DiffusionSat (Khanna et al. 2024). As shown in Figure 1, we augment the frozen, pre-trained U-Net backbone ($\boldsymbol{\epsilon}_\theta$, the blue block) with a trainable, parallel replica of its encoder blocks ($\boldsymbol{\epsilon}_{\theta_c}$, the green block).

This trainable control branch, $\boldsymbol{\epsilon}_{\theta_c}$, exclusively processes the historical image sequence $\mathcal{X}$ to extract a hierarchy of feature maps $\{f_c^{(i)}\}$ extracted from the successive encoder and middle Blocks of the control branch, where i indicates the i^{th} block. To effectively model temporal patterns within these features, each $f_c^{(i)}$ is passed through a novel *Hybrid Temporal Fusion Module (H)*. This module operates with two parallel streams to capture both local and global temporal dependencies. To model local dependencies, one stream uses a 3D convolution to process fine-grained, frame-to-frame changes, resulting in the feature map f_{3D}. In parallel, to capture global dependencies, the other stream uses a temporal attention mechanism (Transformer) to identify long-range patterns across the sequence, producing the feature map f_{attn}. The outputs of these two streams are dynamically fused using a learned, per-block interpolation weight λ_i, allowing the model to adaptively balance its focus on local versus global dynamics:

$$\delta_{\text{ctrl}}^{(i)} = (1 - \lambda_i) \cdot f_{\text{attn}}^{(i)} + \lambda_i \cdot f_{\text{3D}}^{(i)}$$

Crucially, these final control signals $\delta_{\text{ctrl}}^{(i)}$ are then added to the corresponding feature maps of the frozen U-Net's decoder blocks. This additive injection allows *OrgaCast* to guide the generation process with rich spatiotemporal information without disrupting the powerful priors learned by the original backbone.

Side-Channel Conditioning with Metadata and Text To further refine the generation and resolve ambiguity, *OrgaCast* is guided by side-channel conditions derived from metadata $\mathcal{M}$ and text $\mathcal{C}$. Our model processes these conditions using a multi-stream approach, employing two dedicated embedding modules, m_θ and $m_{\theta'}$, for distinct metadata sources. While both modules share an identical architecture, their parameter sets, θ and θ', are independent and not shared. This design enables each module to learn a specialized representation tailored to its respective inputs.

The core architecture of these modules first projects each scalar value (e.g., v_j from the numerical metadata $\mathcal{M} = \{v_1, \dots, v_J\}$) into a high-dimensional space using a sinusoidal positional encoding, $\text{PE}(\cdot)$. To capture feature-specific transformations, these encoded values are then passed through dedicated Multi-Layer Perceptrons (MLPs), each with unique, unshared weights θ_j. The outputs from each stream are aggregated and then summed to form the final metadata embedding:

$$\mathbf{c}_{\text{meta}} = \sum_{j=1}^{J} \text{MLP}_{\theta_j}(\text{PE}(v_j))$$

This embedding $\mathbf{c}_{\text{meta}}$ is added to the diffusion timestep embedding t to provide global, experiment-specific context. For free-text caption $\mathcal{C}$, we use a pre-trained CLIP text encoder τ_θ to produce a semantic embedding $\mathbf{c}_{\text{text}} = \tau_\theta(\mathcal{C})$. This embedding is fed into the cross-attention layers of the U-Net, allowing for fine-grained, spatially-aware textual control.

3.4 Hierarchical Learning Objective

To train *OrgaCast*, we introduce a hierarchical loss function, $\mathcal{L}_{\text{Total}}$, designed to jointly optimize pixel-level accuracy, perceptual realism, and morphological consistency in the predicted fluorescence images $\hat{x}_{T+k}$. This comprehensive objective integrates three complementary components: pixel-wise fidelity ($\mathcal{L}_{\text{MSE}}$), perceptual similarity ($\mathcal{L}_{\text{LPIPS}}$), and morphological coherence ($\mathcal{L}_{\text{Morph}}$).

Pixel-Level Fidelity ($\mathcal{L}_{\text{MSE}}$): The Mean Squared Error (MSE) loss serves as the foundation, enforcing strict pixel-wise correspondence between the generated and ground-truth images:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)^2$$

where n denotes the total number of pixels, and the summation is over all pixel locations i in $\hat{x}_{T+k}$ and x_{T+k}. While MSE ensures basic structural accuracy, it often leads to overly smooth outputs and fails to capture intricate textures characteristic of fluorescence microscopy.

Perceptual Quality ($\mathcal{L}_{\text{LPIPS}}$): To enhance visual realism, we incorporate the Learned Perceptual Image Patch Similarity (LPIPS) loss (Zhang et al. 2018). LPIPS computes perceptual similarity by comparing feature activations from a pretrained deep network:

$$\mathcal{L}_{\text{LPIPS}} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot [\phi_l(\hat{x}_{T+k})_{h,w} - \phi_l(x_{T+k})_{h,w}]\|_2^2$$

where $\phi_l(\cdot)$ represents the feature map from the l-th layer, w_l are learned channel weights, and (h, w) indexes spatial positions. $\mathcal{L}_{\text{LPIPS}}$ encourages the model to reproduce textures and fine details that align with human perceptual judgments.

Morphological Consistency ($\mathcal{L}_{\text{Morph}}$): To ensure predictions respect biologically relevant structures, we introduce a morphological loss that enforces consistency in organoid shape and spatial extent. We define $S(\cdot)$ as a mask extraction function: images are first reduced to single-channel grayscale by summing across fluorescence channels, then thresholded via Otsu's method (Otsu et al. 1975) to obtain binary masks distinguishing foreground organoids. The loss penalizes discrepancies in the predicted and ground-truth foreground areas:

$$\mathcal{L}_{\text{Morph}} = \left| \sum S(\hat{x}_{T+k}) - \sum S(x_{T+k}) \right|$$

This encourages global shape preservation and robustness to morphological variations arising from physical constraints and experimental conditions. Figure 2 provides a visual illustration of the morphological loss in practice.

Total Loss ($\mathcal{L}_{\text{Total}}$): The overall training objective is a weighted sum of the three losses:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MSE}} + \lambda_p \mathcal{L}_{\text{LPIPS}} + \lambda_m \mathcal{L}_{\text{Morph}}$$

where λ_p and λ_m are tunable hyperparameters that balance perceptual and morphological supervision. This hierarchical design ensures that *OrgaCast* produces predictions that are not only accurate at the pixel level, but also visually realistic and biologically faithful—meeting the rigorous demands of downstream scientific analysis.

3.5 Uncertainty-Aware Inference

To provide a reliable measure of predictive confidence under uncertainty, we employ Monte Carlo (MC) Dropout during inference (Gal and Ghahramani 2016). We perform R stochastic forward passes to generate an ensemble of predictions $\{\hat{x}_{T+k}^{(r)}\}_{r=1}^{R}$. The final prediction is the mean of this ensemble:

$$\bar{x}_{T+k} = \frac{1}{R} \sum_{r=1}^{R} \hat{x}_{T+k}^{(r)}$$

We then compute the pixel-wise variance across the ensemble to obtain an uncertainty map U, where high variance indicates low model confidence:

$$U(i,j) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{x}_{T+k}^{(r)}(i,j) - \bar{x}_{T+k}(i,j) \right)^2$$

This map is a crucial tool for expert review, allowing users to identify regions where predictions are less reliable.

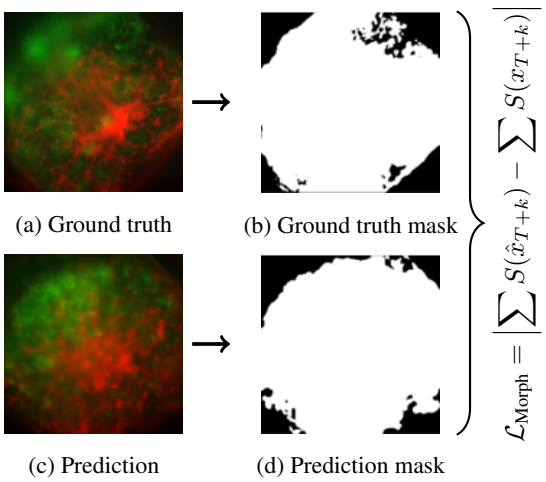


Figure 2: Morphological consistency loss $\mathcal{L}_{\text{Morph}}$ via binary mask area difference between ground truth and prediction.

4 Experiments

We evaluate *OrgaCast* on cardiac organoid forecasting through quantitative and qualitative comparisons with several baselines. We also analyze the model's uncertainty estimates and conduct ablation studies to validate the contribution of each conditioning modality.

4.1 Cardiac Organoid Dataset

Our experiments leverage a time-series fluorescence microscopy dataset capturing the development of vascularized cardiac organoids. The data originates from a genetically engineered hESC-3R cell line designed to express three fluorescent markers, enabling the distinction of key cardiac cell types: TNNT2-GFP (cardiomyocytes, green), CDH5-mOrange (endothelial cells, red), and TAGLN-CFP (smooth muscle cells, blue).

Each cardiac organoid, measuring approximately 1000–1500 µm in diameter, was imaged at five developmental stages (Days 7, 9, 11, 13, and 15), yielding multi-channel image sequences that capture critical spatiotemporal dynamics. To provide essential experimental context, each image sequence is paired with a 4-dimensional metadata vector: [condition_id, is_dynamic, is_dapt, day_number]. This vector encodes the experimental group (`condition_id`), the application of fluid flow (`is_dynamic`), treatment with a Notch signaling inhibitor (`is_dapt`), and the specific imaging day (`day_number`). This combination of rich visual data and structured metadata creates a comprehensive multimodal dataset, ideal for training and evaluating our context-aware forecasting model.

4.2 Experimental Setup and Evaluation Metrics

Our experiments aim to forecast the future state of cardiac organoids at Day 15, conditioned on observations from Days 7, 9, and 11. Each training sample comprises a sequence of 3-channel fluorescence microscopy images $\mathcal{X}$, a structured

metadata vector $\mathcal{M}$ encoding experimental conditions, and a descriptive biological caption $\mathcal{C}$. The model is trained to predict the next-stage image (x_{T+k}) given the inputs $\mathcal{X}$, $\mathcal{M}$, and $\mathcal{C}$. All training and evaluations were conducted on a workstation equipped with two NVIDIA H100 GPUs.

To comprehensively assess prediction quality, we employ three complementary metrics: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). SSIM and PSNR evaluate structural and pixel-wise similarity, with higher scores indicating superior fidelity. LPIPS assesses perceptual similarity in deep feature space, where lower values denote higher perceptual alignment with ground truth. Collectively, they provide a robust evaluation of both low-level accuracy and high-level visual realism.

4.3 Comparison with Baseline Models

To rigorously benchmark *OrgaCast*, we compare its performance against a set of representative baselines spanning generative and spatiotemporal modeling paradigms: Pix2Pix (Isola et al. 2017); TimeSformer (Bertasius, Wang, and Torresani 2021); SRGAN (Ledig et al. 2017); and SwinIR (Liang et al. 2021).

Quantitative Analysis As summarized in Table 1, *OrgaCast* consistently outperforms all baselines across SSIM, PSNR, and LPIPS metrics. It achieves the highest SSIM and PSNR, reflecting superior structural and pixel-wise fidelity, and the lowest LPIPS, indicating closer perceptual alignment with ground truth. These results underscore the advantage of our multimodal diffusion-based approach for biologically meaningful and visually accurate prediction.

Qualitative Analysis Figure 3 presents predicted Day 15 images from all models given fluorescence microscopy sequences from Days 7, 9, and 11. *OrgaCast* yields sharper, more coherent, and biologically plausible outputs than competing methods, with notably improved delineation of cardiomyocyte (green) and endothelial (red) regions, highlighting its strength in modeling fine-grained morphological patterns. In contrast, baseline models often produce blurry artifacts, miss key structural features, or fail to capture the correct organoid morphology.

Model	SSIM ↑	PSNR ↑	LPIPS ↓
Pix2Pix	0.4947	16.9627	0.6685
TimeSformer	0.3478	17.1338	0.7300
SRGAN	0.5047	16.6221	0.5911
SwinIR	0.5477	15.6869	0.5892
***OrgaCast* (Ours)**	**0.5573**	**17.9378**	**0.5483**

Table 1: Comparison of *OrgaCast* against baselines.

4.4 Ablation and Component Analysis

We conduct a series of ablation studies to dissect the contributions of our model's key components, including the multimodal conditioning framework, the hierarchical loss function, and the handling of temporal context.

Impact of Multimodal Conditioning To quantify the benefit of our multimodal design, we systematically ablate each conditioning modality: the spatiotemporal control from historical images ($\mathcal{X}$), the structured metadata ($\mathcal{M}$), and the textual captions ($\mathcal{C}$). The results, summarized in Table 2, demonstrate that the full *OrgaCast* implementation substantially outperforms all reduced variants. As expected, removing the spatiotemporal control module causes the most significant performance degradation, confirming that visual history is the primary predictive signal. However, the removal of metadata and text captions also leads to a clear drop in performance, validating our hypothesis that these side-channel inputs provide crucial contextual cues that refine predictions and resolve ambiguity.

Modality Components			SSIM ↑	PSNR ↑	LPIPS ↓
Captions	Metadata	Images			
	✓	✓	0.5181	16.9759	0.5765
✓		✓	0.5158	17.8845	0.5611
✓	✓		0.4243	14.9905	0.6549
✓	✓	✓	**0.5573**	**17.9378**	**0.5483**

Table 2: Ablation results on conditioning modalities. (✓) denote included modality components.

Impact of Hierarchical Loss Components We validate our hierarchical learning objective by training three model variants: one with only the $\mathcal{L}_{\text{MSE}}$ loss, a second that adds the perceptual $\mathcal{L}_{\text{LPIPS}}$ loss, and our full model which includes the morphological $\mathcal{L}_{\text{Morph}}$ loss. As shown in Table 3, we observe a distinct, incremental improvement with the addition of each component. While adding $\mathcal{L}_{\text{LPIPS}}$ significantly improves perceptual quality (lower LPIPS score), the inclusion of our novel $\mathcal{L}_{\text{Morph}}$ provides a final boost across all metrics. This confirms that a multi-faceted objective that evaluates pixel, perceptual, and structural integrity is essential for generating forecasts that are both visually realistic and biologically coherent.

Loss Components			SSIM ↑	PSNR ↑	LPIPS ↓
$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{LPIPS}}$	$\mathcal{L}_{\text{Morph}}$			
✓			0.5022	17.1970	0.5603
✓	✓		0.5397	17.3376	0.5630
✓		✓	0.5267	17.0875	0.5583
✓	✓	✓	**0.5573**	**17.9378**	**0.5483**

Table 3: Ablation results on the components of our hierarchical loss function $\mathcal{L}_{\text{Total}}$. (✓) denotes included loss functions.

Analysis of Temporal Context Sensitivity A trustworthy scientific model should exhibit predictable and logical behavior. To test this, we analyze *OrgaCast*'s sensitivity to temporal context by training variants where we selectively exclude historical frames from Day 7, 9, or 11. The results

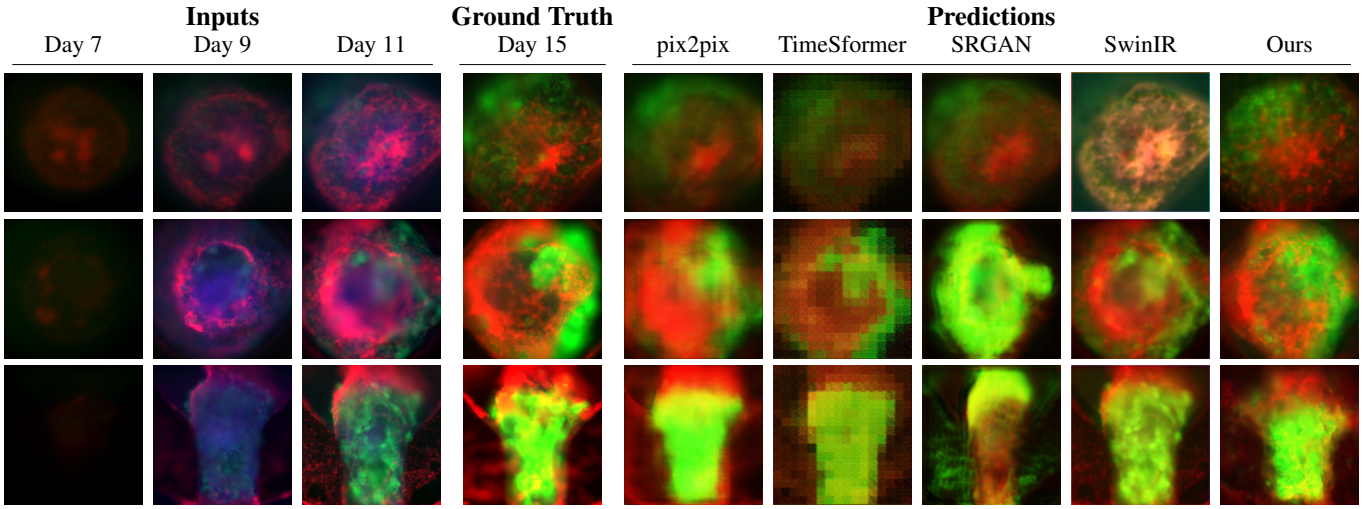


Figure 3: Qualitative comparison across three representative samples. For each row, the columns show Inputs (Days 7, 9, 11), Ground Truth (Day 15), and Predictions generated by pix2pix, TimeSformer, SRGAN, SwinIR, and our method. The diameter of each cardiac organoid ranges from 1000–1500 µm.

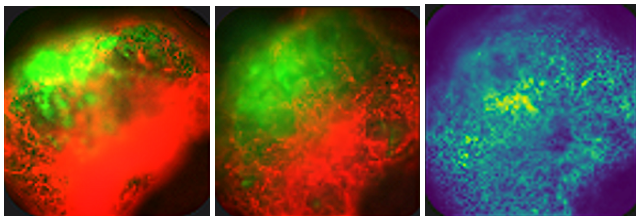
in Table 4 reveal a clear and rational trend: the model's performance is strongly correlated with the temporal proximity of the input frames. Excluding the most recent observation (Day 11) results in a far more significant drop in accuracy than excluding older frames. This logical behavior demonstrates that *OrgaCast* has learned a fundamental principle of forecasting—that recent data is most informative. This predictability is a key component of its design as a reliable and trustworthy scientific instrument, not an opaque black box.

Included Frames			SSIM ↑	PSNR ↑	LPIPS ↓
Day 7	Day 9	Day 11			
✓	✓	✓	**0.5573**	**17.9378**	**0.5483**
	✓	✓	0.5222	16.9426	0.5600
✓		✓	0.5185	16.7299	0.5663
✓	✓		0.4038	13.5283	0.6804

Table 4: Ablation results on the contribution of historical input frames. (✓) denotes the frame was included as input.

4.5 Uncertainty Analysis

A key component of our framework's trustworthiness is its ability to quantify predictive uncertainty. We employ Monte Carlo (MC) Dropout at inference time to generate uncertainty maps, where pixel-wise variance across multiple stochastic forward passes indicates the model's confidence. In these maps (Figure 4), bright yellow signifies high uncertainty, while dark purple denotes high confidence. Our analysis reveals a consistent and informative pattern. The model confidently predicts low-frequency information, such as the overall shape, location, and broad color distribution of the organoid, shown in dark purple. This indicates a robust understanding of the general developmental trajectory. Crucially, the highest uncertainty (bright yellow) is systematically localized within the green fluorescent channel, which corresponds to the TNNT2-GFP for cardiomyocytes. This finding suggests that while *OrgaCast* is confident about the global structure, it correctly identifies the precise texture and intensity of the cardiomyocyte regions, the most dynamic and difficult-to-predict component of cardiac tissue.



(a) Ground truth (b) Prediction (c) Uncertainty map

Figure 4: Uncertainty analysis for a sample prediction.

5 Conclusion

We introduced *OrgaCast*, a novel multimodal conditional diffusion framework for forecasting organoid development from fluorescence microscopy. By synergistically integrating historical images, structured metadata, and text, *OrgaCast* produces high-fidelity, biologically plausible predictions that overcome traditional methods. Comprehensive experiments show *OrgaCast* outperforms strong baselines on pixel-level and perceptual metrics, while ablation and uncertainty analyses validate the multimodal architecture and its ability to capture biological variability. A promising direction is integrating our model into an interactive platform that offers biologists real-time *in silico* experimentation, accelerating discovery and reducing lab costs.

Acknowledgments

This work was supported in part by NIH R15HD108720, NIH R56HL174856, NIH G-RISE T32GM136501, NSF CCF-2447834, and Harry S. Moss Heart Trust.

References

Abilez, O. J.; Yang, H.; Guan, Y.; Shen, M.; Yildirim, Z.; Zhuge, Y.; Venkateshappa, R.; Zhao, S. R.; Gomez, A. H.; El-Mokahal, M.; et al. 2025. Gastruloids enable modeling of the earliest stages of human cardiac and hepatic vascularization. *Science*, 388(6751): eadu9375.

Beilmann, M.; Adkins, K.; Boonen, H. C.; Hewitt, P.; Hu, W.; Mader, R.; Moore, S.; Rana, P.; Steger-Hartmann, T.; Villenave, R.; et al. 2025. Application of new approach methodologies for nonclinical safety assessment of drug candidates. *Nature Reviews Drug Discovery*, 1–21.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *Icml*, volume 2, 4.

Bian, X.; Li, G.; Wang, C.; Liu, W.; Lin, X.; Chen, Z.; Cheung, M.; and Luo, X. 2021. A deep learning model for detection and tracking in high-throughput images of organoid. *Computers in Biology and Medicine*, 134: 104490.

Chen, T.; Zhang, Y.; Xie, L.; Shen, W.; Wu, S.; and Wong, H.-S. 2025. SpotDiff: Spatial Gene Expression Imputation Diffusion with Single-Cell RNA Sequencing Data Integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15848–15856.

de Medeiros, G.; Ortiz, R.; Strnad, P.; Boni, A.; Moos, F.; Repina, N.; Challet Meylan, L.; Maurer, F.; and Liberali, P. 2022. Multiscale light-sheet organoid imaging framework. *Nature Communications*, 13(1): 4864.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Du, X.; Chen, Z.; Li, Q.; Yang, S.; Jiang, L.; Yang, Y.; Li, Y.; and Gu, Z. 2023. Organoids revealed: morphological analysis of the profound next generation in-vitro model with artificial intelligence. *Bio-design and Manufacturing*, 6(3): 319–339.

Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; and Dean, J. 2019. A guide to deep learning in healthcare. *Nature medicine*, 25(1): 24–29.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Hammer, M.; Huisman, M.; Rigano, A.; Boehm, U.; Chambers, J. J.; Gaudreault, N.; North, A. J.; Pimentel, J. A.; Sudar, D.; Bajcsy, P.; et al. 2021. Towards community-driven metadata standards for light microscopy: tiered specifications extending the OME model. *Nature methods*, 18(12): 1427–1440.

He, R.; Sarwal, V.; Qiu, X.; Zhuang, Y.; Zhang, L.; Liu, Y.; and Chiang, J. 2025. Generative AI models in time-varying biomedical data: scoping review. *Journal of Medical Internet Research*, 27: e59792.

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

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kandula, A. K. R.; Phamornratanakun, T.; Gomez, A. H.; El-Mokahal, M.; Ma, Z.; Feng, Y.; and Yang, H. 2025. Generative Ai for Cardiovascular Cell Type-Specific Fluorescence Colorization of Live-Cell hPSC-Derived Cardiac Organoids. *Advanced Intelligent Discovery*, 202400041.

Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D. B.; and Ermon, S. 2024. DiffusionSat: A Generative Foundation Model for Satellite Imagery. In *The Twelfth International Conference on Learning Representations*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning Image Restoration without Clean Data. In *International Conference on Machine Learning*, 2965–2974. PMLR.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27.

Özbey, M.; Dalmaz, O.; Dar, S. U.; Bedel, H. A.; Özturk, Ş.; Güngör, A.; and Cukur, T. 2023. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12): 3524–3539.

Papargyriou, A.; Najajreh, M.; Cook, D. P.; Maurer, C. H.; Bärthel, S.; Messal, H. A.; Ravichandran, S. K.; Richter, T.; Knolle, M.; Metzler, T.; et al. 2024. Heterogeneity-driven phenotypic plasticity and treatment response in branched-organoid models of pancreatic ductal adenocarcinoma. *Nature biomedical engineering*, 1–29.

Pinaya, W. H.; Tudosiu, P.-D.; Dafflon, J.; Da Costa, P. F.; Fernandez, V.; Nachev, P.; Ourselin, S.; and Cardoso, M. J.

2022. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, 117–126. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sermesant, M.; Delingette, H.; Cochet, H.; Jaïs, P.; and Ayache, N. 2021. Applications of artificial intelligence in cardiovascular imaging. *Nature Reviews Cardiology*, 18(8): 600–609.

Tiwari, S. K.; Wang, S.; Smith, D.; Carlin, A. F.; and Rana, T. M. 2021. Revealing tissue-specific SARS-CoV-2 infection and host responses using human stem cell-derived lung and cerebral organoids. *Stem Cell Reports*, 16(3): 437–445.

Van Valen, D. A.; Kudo, T.; Lane, K. M.; Macklin, D. N.; Quach, N. T.; DeFelice, M. M.; Maayan, I.; Tanouchi, Y.; Ashley, E. A.; and Covert, M. W. 2016. Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS computational biology*, 12(11): e1005177.

Wang, A. Q.; Karaman, B. K.; Kim, H.; Rosenthal, J.; Saluja, R.; Young, S. I.; and Sabuncu, M. R. 2024. A framework for interpretability in machine learning for medical imaging. *IEEE Access*, 12: 53277–53292.

Wang, Y.; Long, M.; Wang, J.; Gao, Z.; and Yu, P. S. 2017. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30.

Wolleb, J.; Bieder, F.; Sandkühler, R.; and Cattin, P. C. 2022. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, 35–45. Springer.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Yi, X.; Walia, E.; and Babyn, P. 2019. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58: 101552.

Yilmaz, R.; Eschweiler, D.; and Stegmaier, J. 2024. Annotated biomedical video generation using denoising diffusion probabilistic models and flow fields. In *International Workshop on Simulation and Synthesis in Medical Imaging*, 197–207. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.