EmbryoDiff: A Conditional Diffusion Framework with Multi-Focal Feature Fusion for Fine-Grained Embryo Developmental Stage Recognition

Yong Sun, Zhengjie Zhang, Junyu Shi, Zhiyuan Zhang, Lijiang Liu, Qiang Nie*

The Hong Kong University of Science and Technology (Guangzhou)
qiangnie@hkust-gz.edu.cn

Abstract

Identification of fine-grained embryo developmental stages during In Vitro Fertilization (IVF) is crucial for assessing embryo viability. Although recent deep learning methods have achieved promising accuracy, existing discriminative models fail to utilize the distributional prior of embryonic development to improve accuracy. Moreover, their reliance on single-focal information leads to incomplete embryonic representations, making them susceptible to feature ambiguity under cell occlusions. To address these limitations, we propose EmbryoDiff, a two-stage diffusion-based framework that formulates the task as a conditional sequence denoising process. Specifically, we first train and freeze a frame-level encoder to extract robust multi-focal features. In the second stage, we introduce a Multi-Focal Feature Fusion Strategy that aggregates information across focal planes to construct a 3D-aware morphological representation, effectively alleviating ambiguities arising from cell occlusions. Building on this fused representation, we derive complementary semantic and boundary cues and design a Hybrid Semantic-Boundary Condition Block to inject them into the diffusion-based denoising process, enabling accurate embryonic stage classification. Extensive experiments on two benchmark datasets show that our method achieves state-of-the-art results. Notably, with only a single denoising step, our model obtains the best average test performance, reaching 82.8% and 81.3% accuracy on the two datasets, respectively.

Code — https://github.com/RIL-Lab/EmbryoDiff

Introduction

Infertility has affected approximately 80 million individuals of reproductive age worldwide (Inhorn and Patrizio 2015). In response, in vitro fertilization (IVF) has become a central intervention. In current clinical practice, embryologists examine Time-Lapse Monitoring (TLM) videos of in vitro fertilized embryos to extract both morphological and morphokinetic features, selecting the most viable embryo for transfer, which is labor-intensive and time-consuming. Recently, deep learning has been adopted in assisted reproductive technology, with applications spanning blastocyst evaluation (Kragh et al. 2019; Kromp et al. 2023), embryo grading (Sun et al. 2025), and viability prediction (Kim et al. 2024; Kragh et al. 2021; Borna, Sepehri, and Maleki 2024).

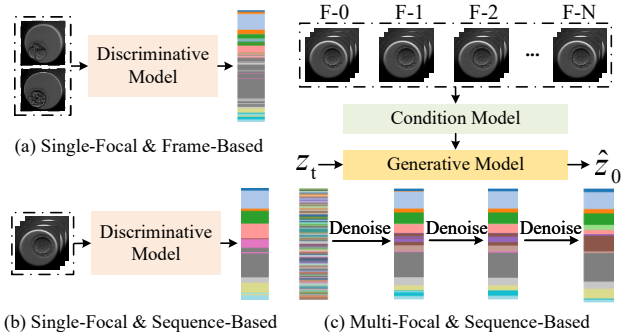

Figure 1: Comparison of different frameworks. (a) Frame-based methods process frames independently. (b) Sequence-based methods capture temporal dependencies from single-view videos. (c) Our method integrates multi-focal TLM videos with a conditional generative model. Different colors indicate different developmental stages.

A crucial subtask in embryo evaluation is embryo developmental stage classification. The goal is to assign a developmental phase to each frame in a TLM sequence, thereby offering morphokinetic cues that are essential for embryo quality assessment (Jacobs et al. 2020; Rienzi et al. 2019). In recent years, considerable effort has been devoted to automating this labor-intensive and expert-dependent process using deep learning techniques. As illustrated in Figure 1(a) and (b), existing approaches can be broadly categorized into two groups: (1) **Frame-based methods** (Liu et al. 2025; Dimitriadis et al. 2019; Dirvanauskas et al. 2019; Guo et al. 2023), which process each frame independently and often suffer from limited temporal coherence due to the absence of temporal modeling; and (2) **Sequence-based methods** (Lukyanenko et al. 2021; Ng et al. 2018; Lockhart et al. 2021; Canat et al. 2024), which take entire TLM sequences as input and explicitly model temporal dependencies among developmental stages. Despite the progress achieved, accurate fine-grained developmental stage classification remains challenging due to several key limitations.

First, existing discriminative methods fail to leverage the inherent distributional priors of embryonic development. Al-

though individual embryos vary in developmental timing, their overall trajectories generally follow well-established biological principles, such as monotonic stage transitions, characteristic phase durations, and the coordinated occurrence of developmental events. Discriminative models, however, learn direct mapping between visual inputs and stage labels without modeling these underlying priors. Consequently, even state-of-the-art approaches are prone to unstable or unreliable predictions under challenging conditions.

Moreover, prior methods suffer from incomplete embryonic morphological representation due to their reliance on single-focal-plane TLM videos, which capture only a partial 2D projection of the embryo's inherently 3D structure. As development progresses through successive cell divisions, the increasing number of blastomeres leads to frequent cell occlusions in these 2D projections. In clinical practice, embryologists mitigate this issue by integrating visual cues across multiple focal planes to reconstruct a more complete 3D morphological context. In contrast, current approaches are restricted to single-focal visual inputs and therefore lack the contextual information required to disambiguate overlapping cells in densely packed configurations.

Furthermore, model performance is constrained by data quality and evaluation practices. Noisy annotations in the public dataset (Barhoun et al. 2025; Gomez et al. 2022) compromise training stability, and frame-level metrics used in current methods ignore the sequence nature of embryo developmental progression.

In this paper, we propose **EmbryoDiff**, a two-stage diffusion-based framework (Figure 1(c)) designed to address these challenges. Unlike discriminative methods, diffusion models learn the underlying data distribution and can therefore naturally encode the developmental priors that govern embryonic progression. Inspired by conditional diffusion architectures such as LDM (Rombach et al. 2022) and DiT (Peebles and Xie 2023), EmbryoDiff formulates the task as a conditional sequence denoising process and uses multi-focal TLM sequences as conditioning inputs for guidance, enabling the recovery of more accurate and biologically plausible developmental stage labels from noisy sequences. To reduce computational burden, we adopt a two-stage training strategy. In the first stage, a frame-based model is trained on individual frames and fixed as a feature extractor. In the second stage, a conditional diffusion model is trained to obtain final classification results. We extract features from multi-focal-plane TLM videos and fuse them via a Multi-Focal Feature Fusion Strategy, constructing more comprehensive, 3D-context-aware morphological representations. To provide comprehensive conditional guidance, we design parallel semantic and boundary condition encoders to capture complementary cues, and introduce a Hybrid Semantic-Boundary Condition Block to effectively inject these features into the denoising process. By leveraging the distributional priors encoded in the diffusion model, enhanced 3D-aware representations, and comprehensive conditional guidance, EmbryoDiff achieves more accurate classification results and outperforms prior methods.

Beyond algorithmic improvements, we also enhance data quality and evaluation protocols. We systematically correct annotations in a public dataset (Gomez et al. 2022) by removing errors and refining stage boundaries. Inspired by action segmentation (Yi, Wen, and Jiang 2021; Lu and Elhamifar 2024), we adopt both frame-level and sequence-level metrics to enable a more comprehensive evaluation.

In summary, our contributions are three-fold:

- We introduce EmbryoDiff, the first diffusion-based framework for embryonic stage recognition, and the first to leverage multi-focal-plane video inputs for holistic embryo representation.
- We enhance the data quality of a public dataset and leverage sequence-level evaluation metrics to complement frame-level accuracy, enabling more reliable training and assessment.
- We conduct extensive experiments and ablations, achieving state-of-the-art results on two benchmark datasets and providing insights for future research.

Related Works

Diffusion Model for Visual Perception. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) generate data samples through an iterative denoising process and have achieved state-of-the-art performance in various generation tasks. Recent studies have shown that, with appropriate conditional guidance, diffusion models can also be effectively applied to visual perception tasks. These models have demonstrated promising results in a wide range of applications, including image segmentation (Wu et al. 2024; Rahman et al. 2023; Gu, Chen, and Xu 2024; Ji et al. 2023), depth estimation (Zhang et al. 2024; Ke et al. 2024), object detection (Chen et al. 2023; Gong, Kwak, and Cho 2024), action segmentation (Liu et al. 2023; Gong, Kwak, and Cho 2024), temporal action localization (Nag et al. 2023) and anomaly detection (Wyatt et al. 2022; He et al. 2024; Zhang et al. 2025). However, in embryo developmental stage recognition, existing methods rely on discriminative models without considering the distribution prior of embryo development. In this paper, we make the first attempt to leverage the diffusion-based framework for this task.

Embryo Developmental Stage Classification. Embryo developmental stage classification provides critical morphokinetic features for embryo selection, such as division timing and stage duration. Recent studies have explored deep learning to automate this process. Early methods adopted CNNs like InceptionV3 (Dimitriadis et al. 2019) and ResNet (Gomez et al. 2022) for frame-wise classification, but ignored temporal dependencies. Later works (Canat et al. 2024; Lukyanenko et al. 2021) incorporated temporal modeling, improving performance. Embryosformer (Nguyen et al. 2023) further enhanced accuracy by integrating deformable attention and transformer decoder. However, accurate identification of fine-grained developmental stages remains challenging by soly using single-focal information. To this end, in this paper, we incorporate multi-focal feature fusion to improve the classification performance.

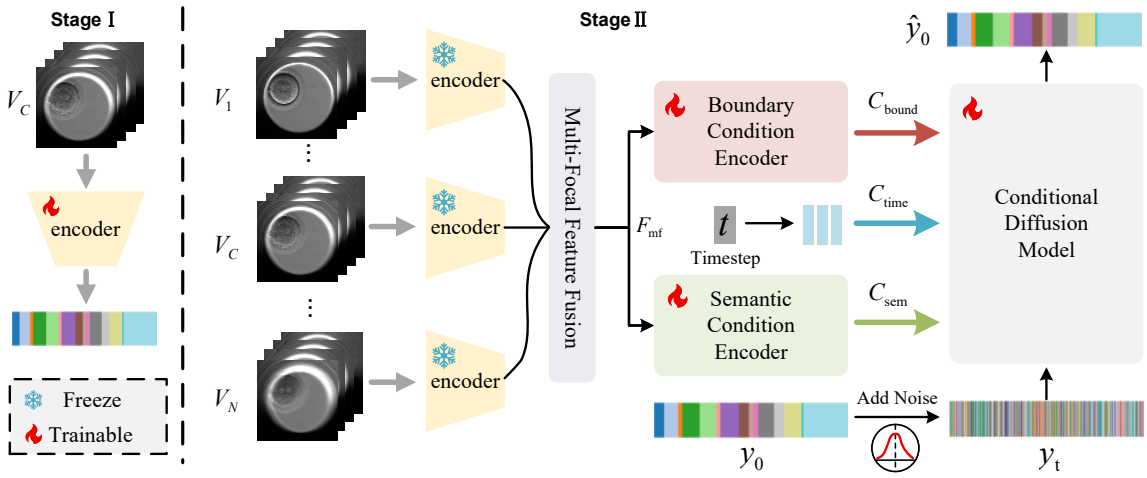

Figure 2: Overview of the two-stage EmbryoDiff framework. Stage 1 (left) trains a frame-based model, while Stage 2 (right) utilizes multi-focal videos to generate semantic and boundary conditions that guide the diffusion-based denoising process.

Method

EmbryoDiff consists of two stages (Figure 2). Stage 1 trains a frame-based model on central-plane TLM videos and freezes it as a multi-focal feature extractor. Stage 2 trains the diffusion model: the extracted multi-focal features are fused by the Multi-Focal Feature Fusion Strategy and subsequently encoded by the Boundary Condition Encoder and the Semantic Condition Encoder to generate robust and complementary conditional signals. These signals are then injected into the diffusion model to recover clean stage sequences from noisy inputs.

Frame-Based Classification Model

Multi-focal TLM videos are 5D data and consist of hundreds of frames per sequence, posing huge computational and memory challenges for end-to-end training. To this end, we adopt a two-stage framework. In Stage 1, following prior work (Gomez et al. 2022), we treat each frame independently and train a frame-based classifier to learn robust visual representations using TLM videos V_c from central focal plane. In practice, we use ResNet-50 (He et al. 2016) as the backbone and optimize it with standard cross-entropy loss.

Diffusion-Based Classification Model

In Stage 2, we condition the diffusion model on TLM sequences to denoise and generate coherent developmental stage predictions from noise corrupted labels.

Training Let $\{V_1, \ldots, V_N\}$ denote the videos from N focal planes. The pretrained frame-level encoder processes each video to extract its feature sequence $F_i = \mathrm{Enc}(V_i) \in \mathbb{R}^{T\times D}$, where $D = 2048$. These features are then fused by the Multi-Focal Feature Fusion Strategy to form a holistic representation $F_{\mathrm{mf}} = \Phi(F_1, \ldots, F_N)$, where $\Phi(\cdot)$ denotes the fusion operation. The fused representation is further encoded by the Semantic Condition Encoder and the Boundary Condition Encoder to obtain the complementaru semantic condition C_{sem} and boundary condition C_{bound}. For training the diffusion model, inspired by diffusion-based segmentation methods (Ji et al. 2023; Wu et al. 2024; Liu et al. 2023), we add noise to ground-truth labels and train the diffusion model to reverse this process and recover clean sequences. Suppose that we have c development stages. We first map the T-length one-hot stage labels $\mathbf{y} \in \mathbb{R}^{T\times c}$ into semantic embeddings $\mathbf{y}_0 \in \mathbb{R}^{T\times d}$ using a learnable embedding layer, where d denotes the embedding dimension. We then apply the standard forward diffusion process to corrupt the label sequence with Gaussian noise:

$$q(\mathbf{y}_t \mid \mathbf{y}_0) = \mathcal{N}(\mathbf{y}_t; \sqrt{\bar{\alpha}_t}\mathbf{y}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative noise scaling factor at time step t. To reverse the process, the diffusion model receives the noisy sequence $\mathbf{y}_t$, the timestep embedding C_{time}, and two types of conditional features C_{sem} and C_{bound} to make predictions:

$$\hat{\mathbf{y}}_0 = f_\theta(\mathbf{y}_t, C_{\mathrm{time}}, C_{\mathrm{sem}}, C_{\mathrm{bound}}), \tag{2}$$

where f_θ denotes the diffusion model parameterized by θ. We adopt a DiT-style (Peebles and Xie 2023) diffusion decoder, where the timestep embedding C_{time} is first added to the noisy features and then processed by stacked Hybrid Semantic-Boundary Condition Blocks to inject the two complementary condition signals into the denoising process. Finally, the predicted embeddings $\hat{\mathbf{y}}_0$ are mapped back to the one-hot space to obtain the predicted labels $\hat{\mathbf{y}}$.

Inference In the inference phase, we sample a random noise from the standard Gaussian distribution.

$$\mathbf{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

Then, the DDIM (Song, Meng, and Ermon 2020) sampling strategy is utilized and we can generate high-quality predictions through several denoising steps.

Multi-Focal Feature Fusion

Embryonic development is a dynamic process of cell division and 3D reorganization, often causing cell occlusions

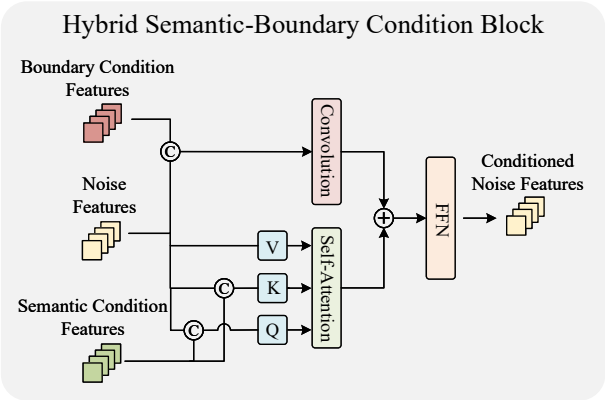

Figure 3: The architecture of the Hybrid Semantic-Boundary Condition Block. For clarity, the residual connection between the original and conditioned noise features is omitted.

that hinder feature learning (Lukyanenko et al. 2021). To capture a more comprehensive morphology, we integrate features from multiple focal planes. Specifically, we adopt a prefusion strategy, in which features from multi-focal videos are fused before entering the condition encoders. This design brings two key benefits. First, it substantially reduces the computational cost of the encoders by avoiding attention over excessively long feature sequences. Second, it improves the framework's flexibility, allowing it to seamlessly accommodate single-focal settings by simply disabling the fusion step without modifying any other components. In practice, we employ a simple average pooling operation along the focal-plane dimension to obtain the fused feature $F_{\text{mf}} \in \mathbb{R}^{T \times D}$ as:

$$F_{\text{mf}} = \frac{1}{N} \sum_{i=1}^{N} F_i, \tag{4}$$

This strategy shifts from 2D in-plane analysis to a 3D-aware modeling of embryonic structure, capturing spatial context that enhances feature robustness against cell occlusions.

Complementary Semantic-Boundary Conditions

The performance of conditional diffusion models is highly dependent on the quality of conditioning signals. Prior work MedSefDiff-V2 (Wu et al. 2024) uses an auxiliary segmentation loss to train the condition network for semantic guidance in denoising. However, such semantic cues alone provide limited discriminative power for embryo stage recognition, particularly in transitional phases where morphological changes are subtle and stage boundaries are difficult to distinguish. To this end, we introduce complementary boundary-aware condition features that work in conjunction with semantic cues to provide more discriminative guidance during the denoising process. As shown in Figure 2, a dual-branch encoder extracts semantic features C_{sem} and boundary features C_{bound} from the fused multi-focal sequence features F_{mf}. Both branches adopt the ASFormer architecture (Yi, Wen, and Jiang 2021), capturing local and global temporal dependencies.

Hybrid Semantic-Boundary Condition Block As shown in Figure 3, each block handles noise and condition features via two complementary pathways. In the boundary-aware path, boundary condition features are concatenated with noise features and processed by a convolution layer to capture local interactions. In the semantic-aware path, semantic condition features are concatenated with noise features to form the query-key pairs, while the noise features serve as values in a self-attention module to inject long-range semantic context. The outputs from both pathways are summed and passed through a Feed-Forward Network (FFN) to produce the final conditioned noise features.

This hybrid design ensures the synergistic integration of local boundary cues and global semantic context, assisting the denoising process to generate both semantic correct and boundary-accurate predictions.

Loss Functions

Multiple losses are jointly optimized in Stage 2. The semantic branch uses a stage classification loss $\mathcal{L}_{\text{sem}}$ and a temporal smoothing loss $\mathcal{L}_{\text{smooth}}$ (Yi, Wen, and Jiang 2021), while the boundary branch uses a binary boundary classification loss $\mathcal{L}_{\text{bound}}$. The diffusion model is also optimized with a stage classification loss $\mathcal{L}_{\text{diff}}$. The overall objective is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{sem}} + \lambda_2 \mathcal{L}_{\text{smooth}} + \lambda_3 \mathcal{L}_{\text{bound}} + \lambda_4 \mathcal{L}_{\text{diff}}, \tag{5}$$

All classification losses are standard cross-entropy losses. We set λ_1=0.8, λ_2=0.3, λ_3=0.5, and λ_4=1.0 to balance their contributions.

Experiments

Datasets

Multi-Focal Human Embryos Dataset (MFHE) We use the MFHE dataset (Gomez et al. 2022) to evaluate the effectiveness of multi-focal feature fusion and the proposed diffusion framework. It contains 704 embryo sequences with 7 focal planes and 16 developmental stages. We refine the annotations and remove the rare tHB class, resulting in 655 videos. The dataset is split 7:3 into train and test sets.

Single-Focal Human Embryos Dataset (SFHE) We use the SFHE dataset (Nguyen et al. 2023) to assess the generality and single-focal adaptability of our framework. It contains 440 embryo sequences with single-focal videos and 12-stage labels. We use the original train set for training and merge the validation and test sets for testing.

Evaluation Metrics

To evaluate temporal consistency and boundary localization accuracy, we use standard metrics from action segmentation: Accuracy, Edit Score, and F1@10, 25, 50. While Accuracy measures frame-wise correctness, Edit Score and F1 scores assess temporal segment quality, offering a more comprehensive evaluation.

Implementation Details

In Stage 1, a ResNet-50 (He et al. 2016) backbone is trained using the AdamW optimizer, a batch size of 32, and an initial

	Methods	Multi-Focal Human Embryos Dataset				Single-Focal Human Embryos Dataset			
		Acc	Edit	F1@{10, 25, 50}	Avg	Acc	Edit	F1@{10, 25, 50}	Avg
Frame	ResNet	74.3	24.7	30.3 / 26.3 /19.1	34.9	75.8	24.1	30.4 / 26.1 / 18.9	35.1
	DLTEmbryo	69.2	24.3	28.4 / 24.3 / 17.8	32.8	71.3	22.5	27.3 / 22.7 / 15.8	31.9
	InceptionV3	76.2	29.8	34.7 / 30.9 / 23.9	39.1	75.9	27.0	34.1 / 29.5 / 22.1	37.7
Sequence	ResNet-LSTM	75.0	46.5	51.2 / 45.7 / 35.7	50.8	75.6	46.4	54.9 / 49.0 / 37.3	52.6
	LateFusion	77.6	64.2	66.3 / 60.6 / 50.3	63.8	76.1	60.2	66.1 / 59.7 / 45.6	61.5
	ASFormer	77.7	**89.0**	78.7 / 73.0 / 59.1	75.5	76.6	**90.5**	87.4 / 81.3 / 64.0	80.0
	EmbryosFormer	77.4	71.0	72.6 / 69.4 / 61.5	70.4	77.0	81.0	81.3 / 75.0 / 61.6	75.2
	FACT	78.9	83.6	79.6 / 74.5 / 65.6	76.4	76.1	87.0	86.5 / 80.8 / 65.9	79.3
EmbryoDiff (1 step)		82.8	82.8	81.0 / 75.9 / 65.8	77.7	81.3	83.5	85.9 / 81.1 / 68.9	80.1
EmbryoDiff (15 steps)		83.1	86.3	83.4 / 78.2 / 68.4	79.9	81.2	86.8	88.1 / 83.7 / 71.2	82.2
EmbryoDiff (25 steps)		**83.1**	86.7	**83.7 / 78.6 / 68.9**	**80.2**	**81.3**	86.8	**88.4 / 84.2 / 71.3**	**82.4**

Table 1: Comparison with the state-of-the-art methods on Multi-Focal Human Embryos Dataset and Single-Focal Human Embryos Dataset. In the table, "Frame" means frame-based methods and "Sequence" stands for sequence-based methods.

Methods	tPB2	tPNa	tPNf	t2	t3	t4	t5	t6	t7	t8	t9+	tM	tSB	tB	tEB
ResNet	64.4	92.0	91.7	90.6	49.3	75.6	**38.2**	27.0	26.7	65.6	75.3	75.4	66.9	49.3	88.2
DLTEmbryo	34.6	91.3	76.7	82.2	31.1	78.7	11.8	15.2	4.5	76.0	68.0	70.5	65.9	24.9	91.0
InceptionV3	63.2	94.7	89.5	90.2	**55.9**	80.5	26.7	23.1	27.4	73.5	80.1	72.5	65.3	42.9	90.1
ResNet-LSTM	58.9	95.0	90.3	92.5	53.4	78.6	35.1	31.6	**30.2**	61.9	77.7	66.6	71.8	34.2	**91.2**
LateFusion	91.9	95.1	91.3	95.1	47.6	79.1	36.0	31.2	34.1	61.0	78.8	73.2	72.9	48.4	91.1
ASFormer	87.6	96.2	77.5	90.7	28.6	86.6	20.5	28.5	22.5	78.6	74.3	73.0	78.9	37.4	93.4
EmbryosFormer	86.0	97.1	74.2	95.7	0.0	**97.4**	0.7	8.3	0.0	**86.1**	81.1	58.3	**87.2**	2.5	84.9
FACT	88.6	95.8	83.0	96.0	17.2	87.2	28.3	22.2	23.2	75.3	80.1	75.2	75.2	51.4	90.1
EmbryoDiff	**93.2**	**97.2**	**93.7**	**97.0**	51.1	90.9	36.0	**35.0**	28.5	83.4	**81.8**	**79.0**	84.9	**52.7**	90.4

Table 2: Per-class accuracy comparison on MFHE Dataset. For EmbryoDiff, we use 25 denoising steps.

learning rate of 1×10^{-4} with cosine decay over 20 epochs. In Stage 2, frame-wise features with dimension $D = 2048$ are extracted from all focal planes. Each condition encoder consists of 6 layers with a reduced hidden dimension of 96. Intermediate features from $\{2, 4, 6\}$ layers are concatenated to form the condition signals. The diffusion model consists of 8 hybrid conditioning blocks with a feature dimension of 128. We follow the cosine noise schedule with 1000 diffusion steps and set the input signal scale to 0.1. The model is trained for 350 epochs with a batch size of 24, an initial learning rate of 1×10^{-4}, and a weight decay of 0.01 using AdamW and cosine learning rate decay. All models are trained on the training set using a single NVIDIA RTX 4090 GPU and the best evaluation results on test set are reported. For more details about the dataset and experimental setups, please see the online extended version.

Comparison with SOTAs

We compare EmbryoDiff with frame-based (e.g., ResNet (Gomez et al. 2022)), sequence-based (e.g., EmbryosFormer (Nguyen et al. 2023)), and top-performing action segmentation models (ASFormer (Yi, Wen, and Jiang 2021), FACT (Lu and Elhamifar 2024)). Open-source methods are evaluated using official codes; closed-source ones (DLTEmbryo (Liu et al. 2025), LateFusion (Ng et al. 2018)) are carefully reimplemented. For our model, we report average results over ten runs with different random seeds.

Overall Performance Comparison Detailed results are listed in Table 1. On MFHE dataset, frame-based methods such as InceptionV3 achieve reasonable frame-level accuracy (76.2%) but suffer in sequence-level metrics (F1@50: 23.9, Edit: 29.8) due to lack of temporal modeling. In contrast, sequence-based models like FACT, which uses complex designs to model temporal dependencies, achieve better accuracy (78.9%) and significantly improved F1@50 (65.6%). However, they still struggle with challenging samples, likely due to their discriminative nature and reliance on single-focal videos. In comparison, our method outperforms all baseline methods on the average metric with only a single denoising step. As the denoising steps increases to 25, EmbryoDiff achieves 83.1% Accuracy and 68.9% F1@50, setting a new state of the art. While ASFormer performs well on the Edit score (89.0), it lags behind our model on other metrics, particularly F1@50 (59.1%), indicating its limited ability to precisely localize stage boundaries.

On SFHE dataset, where all methods use single-focal input, our EmbryoDiff approach still achieves the best average metric using only one denoising step, demonstrating

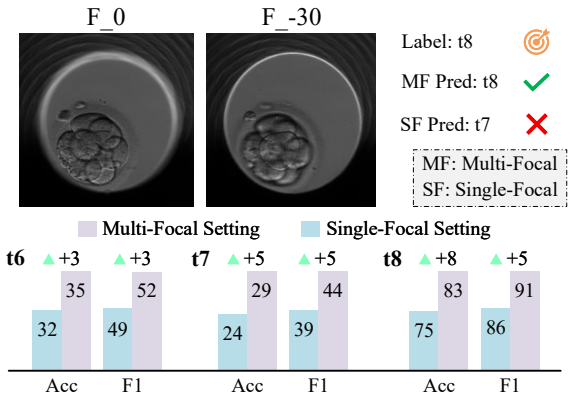

Figure 4: Advantages of multi-focal feature fusion over single-focal. The top shows it better handles cell occlusions, and the bottom shows it greatly improves late cleavage stage identification.

the superiority of our diffusion-based framework. With 25 sampling steps, our method reaches 81.3% Accuracy, improving upon the second-best method, EmbryosFormer, by 5.58%. Moreover, our F1@50 reaches 71.3%, surpassing the second-best FACT (65.9%) by 8.19%.

Per-Class Accuracy Comparison In Table 2, we present a per-class accuracy comparison across methods on MFHE dataset. Our proposed EmbryoDiff achieves the best or second-best accuracy on the majority of classes, demonstrating its strong discriminative capability across diverse developmental stages. Notably, for the multi-cell stages (t6, t7, t8, t9+), EmbryoDiff consistently delivers top-tier performance. In contrast, while some methods may excel on one stage, they often suffer from severe performance drops on others. For example, EmbryosFormer achieves 86.1% accuracy on t8, but performs poorly on other stages, with only 8.3% on t6 and 0.0% on t7. The results highlight the superiority of our method in fine-grained embryo stage recognition.

Ablation Studies

Effectiveness of Key Designs In Table 3, we ablate key components of EmbryoDiff on MFHE dataset. The designs include Multi-Focal Feature Fusion (MFF), complementary Semantic Condition Encoder (SCE) and Boundary Condition Encoder (BCE), and the Conditional Diffusion Model (CDM). Specifically, comparing Rows 1 and 2 as well as Rows 3 and 4 reveals that both the discriminative model (SCE only) and the diffusion-based generative model benefit from MFF, with the latter showing a more significant improvement. This highlights the enhanced capacity of the diffusion framework to leverage multi-focal information. Additionally, the diffusion-based framework outperforms the semantic encoder-only architecture (Row 1 vs. Row 3 and Row 2 vs. Row 4), demonstrating its strength in leveraging the distribution priors of embryo development to improve the predictions. Finally, the comparison across Rows 4–6 confirms that both the proposed complementary BCE and MFF contribute substantially to performance gains. Each

SCE	CDM	BCE	MFF	Acc	Edit	F1@50
✓				80.1	78.3	60.2
✓			✓	81.3	81.2	63.0
✓	✓			81.0	82.8	62.2
✓	✓		✓	82.5	85.3	66.0
✓	✓	✓		80.6	83.3	63.9
✓	✓	✓	✓	**83.1**	**86.3**	**68.4**

Table 3: Ablation study results of key designs on MFHE dataset using 15 denoising steps.

SCE	CDM	BCE	Acc	Edit	F1@50
✓			80.5	81.8	67.5
✓	✓		80.8	85.8	70.2
✓	✓	✓	**81.2**	**86.8**	**71.2**

Table 4: Ablation study results of key designs on SFHE dataset using 15 denoising steps.

component consistently improves prediction accuracy and sequence-level metrics, confirming our statements.

The ablation results on the SFHE dataset (Table 4) exhibit similar trends. The conditional diffusion model significantly outperforms the baseline that uses only semantic condition encoder. By incorporating boundary condition features through the proposed Hybrid Semantic-Boundary Condition Block, the prediction performance improves further, particularly on sequence-level metrics such as Edit Score and F1@50. This indicates that boundary-aware signals help the diffusion model better localize transitional frames in the developmental sequence.

Superiority of Multi-Focal Feature Fusion To highlight the critical role of multi-focal feature fusion in the task of fine-grained embryo developmental stage recognition, we perform an in-depth analysis in Figure 4. As shown in the upper part, the central focal plane image (left) clearly reveals seven cells, along with a faint, out-of-focus contour beneath. Relying solely on a single focal plane may lead to misinterpretation of cell count and, consequently, incorrect stage classification. In contrast, the TLM image at a deeper focal plane (right) unambiguously reveals an additional large cell, confirming its presence. By integrating information across multiple focal planes, our method reconstructs a more accurate 3D morphological structure, enabling correct developmental stage assignment.

In the lower part of Figure 4, we compare prediction results for late cleavage multi-cellular stages under single-focal and multi-focal settings. These stages are particularly challenging due to frequent cell occlusions. As shown, multi-focal fusion yields significant performance gains in the t6, t7, and t8 stages compared to the single-focal counterpart. Moreover, the advantage becomes increasingly pronounced as the cell count increases, demonstrating its effectiveness in resolving structural ambiguities in densely packed and occluded configurations.

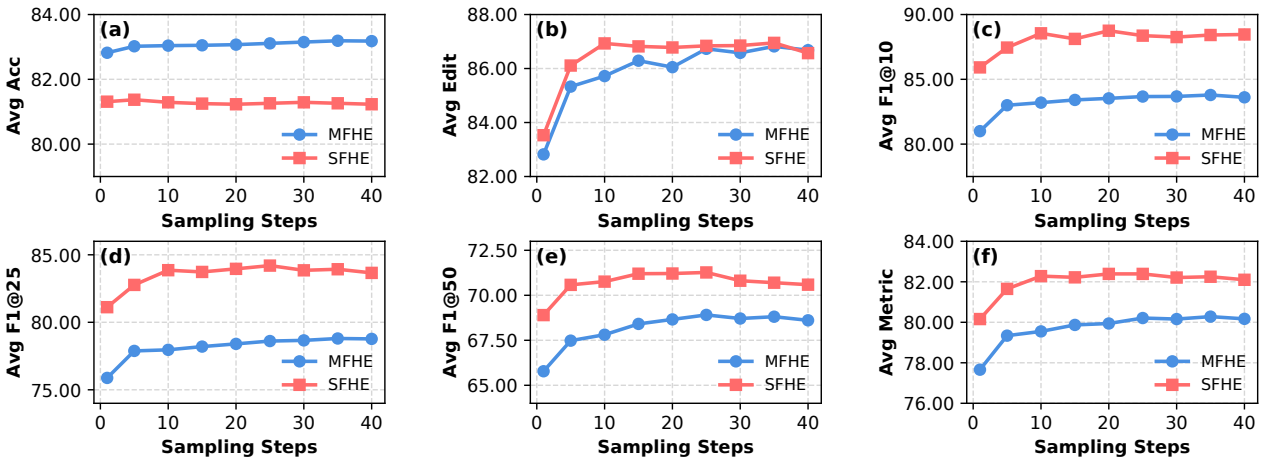

Figure 5: The trends of evaluation metrics on two datasets (MFHE and SFHE) with respect to denoising sampling steps during inference. From (a) to (f), the metrics are Accuracy, Edict Score, F1@{10, 25, 50}, and Average Metric, respectively.

Planes	Acc	Edit	F1@{10, 25, 50}	Avg
1	80.6	83.3	80.4/74.9/63.9	76.6
3	81.4	84.8	81.9/76.7/66.1	78.2
5	82.5	86.2	82.7/77.8/67.9	79.4
7	**83.1**	**86.3**	**83.4/78.2/68.4**	**79.9**

Table 5: Ablation study results of the number of focal planes used with 15 denoising steps. We symmetrically select focal planes around the central focal plane.

Effect of Focal Plane Numbers In Table 5, we investigate the impact of fusing varying numbers of focal planes on model performance. It is evident that prediction accuracy improves significantly as more focal plane information is incorporated. This indicates that multi-focal fusion helps the model better capture 3D embryonic structure and improve recognition performance.

Effect of Denoising Steps We further analyze the impact of inference-time denoising steps on model performance. As shown in Figure 5, the sequence-level metrics (b–f) improve markedly as the number of diffusion steps increases, suggesting that the model progressively refines its predictions by leveraging learned developmental distribution priors. Performance plateaus after about 20 steps. In contrast, accuracy (a) shows only marginal gains. We hypothesize that it is limited by the bottleneck of the condition features.

Qualitative Comparison

In Figure 6, we visualize predicted developmental stage sequences from several leading models. Frame-based methods (e.g., InceptionV3), which lack explicit temporal modeling, produce temporally inconsistent predictions. Sequence-based approaches achieve better temporal coherence but still exhibit significant classification errors and inaccurate boundary localization. In contrast, our model achieves more accurate boundary detection and generates sequences that

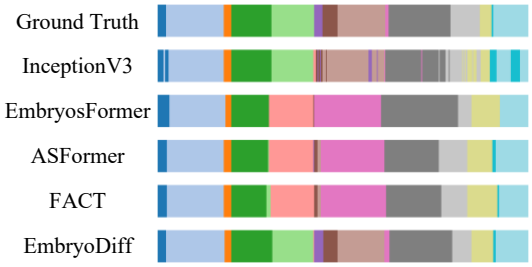

Figure 6: An example of qualitative comparison of developmental stage predictions on the MFHE dataset.

are more consistent with embryonic development patterns, demonstrating its superiority. Additional results are provided in the online extended version.

Conclusion

In this paper, we propose EmbryoDiff, the first diffusion-based framework for embryonic stage recognition, introducing a novel two-stage paradigm that leverages conditional diffusion models to generate accurate and temporally coherent stage sequences. By incorporating TLM video features as conditional inputs, our model guides the denoising process to recover clean stage labels from noisy predictions. To address the issue of incomplete morphological representation, we innovatively fuse multi-focal-plane TLM features, capturing a more holistic view of the 3D embryo structure. Furthermore, to provide richer conditional guidance, we extract complementary semantic and boundary cues from the fused sequence features and design a Hybrid Semantic-Boundary Condition Block to effectively integrate them into the diffusion process. Finally, we improve the annotation quality of an existing dataset and adopt more comprehensive evaluation metrics. Extensive experiments and visualizations demonstrate the superiority of our method.

References

Barhoun, A.; Balafar, M. A.; Oskouei, A. G.; and Sadeghi, L. 2025. Human embryo stage classification using an enhanced R (2+ 1) D model and dynamic programming with optimized datasets. *Biomedical Signal Processing and Control*, 107: 107841.

Borna, M.-R.; Sepehri, M. M.; and Maleki, B. 2024. An artificial intelligence algorithm to select most viable embryos considering current process in IVF labs. *Frontiers in artificial intelligence*, 7: 1375474.

Canat, G.; Duval, A.; Gidel-Dissler, N.; and Boussommier-Calleja, A. 2024. A novel deep learning approach to identify embryo morphokinetics in multiple time lapse systems. *Scientific Reports*, 14(1): 29016.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Dimitriadis, I.; Bormann, C.; Thirumalaraju, P.; Kanakasabapathy, M.; Gupta, R.; Pooniwala, R.; Souter, I.; Hsu, J.; Rice, S.; Bhowmick, P.; et al. 2019. Artificial intelligence-enabled system for embryo classification and selection based on image analysis. *Fertility and sterility*, 111(4): e21.

Dirvanauskas, D.; Maskeliunas, R.; Raudonis, V.; and Damasevicius, R. 2019. Embryo development stage prediction algorithm for automated time lapse incubators. *Computer methods and programs in biomedicine*, 177: 161–174.

Gomez, T.; Feyeux, M.; Normand, N.; David, L.; Paul-Gilloteaux, P.; Fréour, T.; and Mouchère, H. 2022. Towards deep learning-powered IVF: a large public benchmark for morphokinetic parameter prediction. *arXiv preprint arXiv:2203.00531*.

Gong, D.; Kwak, S.; and Cho, M. 2024. Actfusion: a unified diffusion model for action segmentation and anticipation. *Advances in Neural Information Processing Systems*, 37: 89913–89942.

Gu, Z.; Chen, H.; and Xu, Z. 2024. Diffusioninst: Diffusion model for instance segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2730–2734. IEEE.

Guo, W.; Liu, S.; Gong, Z.; Zhang, G.; and Jiang, X. 2023. Cascaded networks for the embryo classification on microscopic images using the residual external-attention. *International journal of imaging systems and technology*, 33(1): 312–322.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; and Xie, L. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8472–8480.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Inhorn, M. C.; and Patrizio, P. 2015. Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. *Human reproduction update*, 21(4): 411–426.

Jacobs, C.; Nicolielo, M.; Erberelli, R.; Mendez, F.; Fanelli, M.; Cremonesi, L.; Aiello, B.; and Lorenzon, A. R. 2020. Correlation between morphokinetic parameters and standard morphological assessment: what can we predict from early embryo development? A time-lapse-based experiment with 2085 blastocysts. *JBRA Assisted Reproduction*, 24(3): 273.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21741–21752.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9492–9502.

Kim, J.; Shi, Z.; Jeong, D.; Knittel, J.; Yang, H. Y.; Song, Y.; Li, W.; Li, Y.; Ben-Yosef, D.; Needleman, D.; et al. 2024. Multimodal learning for embryo viability prediction in clinical ivf. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 542–552. Springer.

Kragh, M. F.; Rimestad, J.; Berntsen, J.; and Karstoft, H. 2019. Automatic grading of human blastocysts from time-lapse imaging. *Computers in biology and medicine*, 115: 103494.

Kragh, M. F.; Rimestad, J.; Lassen, J. T.; Berntsen, J.; and Karstoft, H. 2021. Predicting embryo viability based on self-supervised alignment of time-lapse videos. *IEEE Transactions on Medical Imaging*, 41(2): 465–475.

Kromp, F.; Wagner, R.; Balaban, B.; Cottin, V.; Cuevas-Saiz, I.; Schachner, C.; Fancsovits, P.; Fawzy, M.; Fischer, L.; Findikli, N.; et al. 2023. An annotated human blastocyst dataset to benchmark deep learning architectures for in vitro fertilization. *Scientific data*, 10(1): 271.

Liu, D.; Li, Q.; Dinh, A.-D.; Jiang, T.; Shah, M.; and Xu, C. 2023. Diffusion action segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10139–10149.

Liu, X.; Yu, M.; Liu, H.; Ma, C.; Du, W.; Wu, H.; and Zhang, Y. 2025. DLT-Embryo: A Dual-branch Local feature fusion enhanced Transformer for Embryo multi-stage classification. *Biomedical Signal Processing and Control*, 102: 107266.

Lockhart, L.; Saeedi, P.; Au, J.; and Havelock, J. 2021. Automating embryo development stage detection in time-lapse imaging with synergic loss and temporal learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 540–549. Springer.

Lu, Z.; and Elhamifar, E. 2024. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18175–18185.

Lukyanenko, S.; Jang, W.-D.; Wei, D.; Struyven, R.; Kim, Y.; Leahy, B.; Yang, H.; Rush, A.; Ben-Yosef, D.; Needleman, D.; et al. 2021. Developmental stage classification of embryos using two-stream neural network with linear-chain conditional random field. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 363–372. Springer.

Nag, S.; Zhu, X.; Deng, J.; Song, Y.-Z.; and Xiang, T. 2023. Difftad: Temporal action detection with proposal denoising diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10362–10374.

Ng, N. H.; McAuley, J.; Gingold, J. A.; Desai, N.; and Lipton, Z. C. 2018. Predicting Embryo Morphokinetics in Videos with Late Fusion Nets & Dynamic Decoders.

Nguyen, T.-P.; Pham, T.-T.; Nguyen, T.; Le, H.; Nguyen, D.; Lam, H.; Nguyen, P.; Fowler, J.; Tran, M.-T.; and Le, N. 2023. Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1981–1990.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Rahman, A.; Valanarasu, J. M. J.; Hacihaliloglu, I.; and Patel, V. M. 2023. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11536–11546.

Rienzi, L.; Cimadomo, D.; Delgado, A.; Minasi, M. G.; Fabozzi, G.; Del Gallego, R.; Stoppa, M.; Bellver, J.; Giancani, A.; Esbert, M.; et al. 2019. Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study. *Fertility and sterility*, 112(6): 1080–1093.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, Y.; Wang, Y.; Shi, J.; Zhang, Z.; Xiao, Y.; Zhu, L.; Jiang, M.; and Nie, Q. 2025. Time-Lapse Video-Based Embryo Grading via Complementary Spatial-Temporal Pattern Mining. *arXiv preprint arXiv:2506.04950*.

Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; and Xu, Y. 2024. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, 1623–1639. PMLR.

Wyatt, J.; Leach, A.; Schmon, S. M.; and Willcocks, C. G. 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 650–656.

Yi, F.; Wen, H.; and Jiang, T. 2021. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*.

Zhang, F.; You, S.; Li, Y.; and Fu, Y. 2024. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11852–11861.

Zhang, H.; Wang, Z.; Zeng, D.; Wu, Z.; and Jiang, Y.-G. 2025. DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.