Artificial Intelligence in the Life Sciences 9 (2026) 100170

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

EmbryoTempoFormer: clip-based developmental tempo inference from zebrafish brightfield time-lapse microscopy

Li-Jia-Yu Deng [a,*], Pei-Ran Lin [b], Luo-Tong Xie [c]

[a] Research School of Biology, Australian National University, Canberra ACT 2601, Australia
[b] School of Life Sciences, Westlake University, Hangzhou 310030, PR China
[c] SDU-ANU Joint Science College, Shandong University, Weihai 264209, PR China

ARTICLE INFO

Keywords:
Developmental tempo
Zebrafish embryo staging
Time-lapse microscopy
Temporal context modeling
Embryo-resolved inference

ABSTRACT

Nominal hours post fertilization (hpf) are widely used to index zebrafish embryogenesis, but under condition shifts such as temperature change, genetic perturbation, or environmental stress, nominal time can decouple from true developmental progression. In these settings, biologically meaningful variation is better understood as a change in developmental tempo rather than as a simple temporal offset. Here, we introduce an embryo-resolved framework for developmental-tempo inference from brightfield time-lapse imaging. We present EmbryoTempoFormer (ETF), a clip-based CNN–Transformer that uses short temporal context to predict developmental progression and incorporates a within-embryo temporal-difference consistency regularizer to improve trajectory coherence. We further couple model prediction to an embryo-level inference workflow, in which clip-level outputs are aggregated into interpretable embryo-level tempo and stability summaries, and cross-condition effects are quantified using embryo-bootstrap confidence intervals with embryos—rather than frames or clips—as the independent units. Using a matched two-condition temperature comparison (28.5 °C versus 25 °C) as a proof-of-concept, we show that developmental delay is captured primarily as a reduction in developmental tempo and can be quantified robustly at the embryo level. Together, ETF provides a statistically principled framework for embryo-resolved developmental-tempo analysis in brightfield time-lapse microscopy, demonstrated here in a controlled temperature-shift setting rather than across a broader perturbation panel.

1. Introduction

Zebrafish (*Danio rerio*) are widely used in developmental biology, genetics, toxicology, and drug screening because of their transparent embryos, rapid development, and compatibility with high-throughput perturbation experiments. Accurate staging is fundamental for aligning phenotypes across time and experimental conditions. Traditionally, staging combines nominal hours post fertilization (hpf) with morphological criteria summarized in standard staging guides [1]. In practice, however, investigators are often concerned not only with whether a specific stage has been reached, but also with whether genetic perturbations, drug treatments, or environmental changes alter developmental progression, by how much, and in what way across embryos at the population level. Manual staging is therefore labor-intensive, difficult to scale, and subject to inter-observer variability, limiting stable quantification on large-scale imaging data.

A further challenge is that nominal hpf is not a universal measure of developmental progress. Even under standard temperature (28.5 °C), hpf is only an approximation; under condition shifts—most notably temperature—developmental dynamics can change systematically, and progression may vary across embryos and conditions, making nominal time an imperfect proxy for developmental state [2,3]. In such settings, biologically meaningful variation is often better understood as a change in developmental tempo rather than as a simple additive shift in time. Developmental delay may therefore reflect altered speed and rhythm of progression, potentially in stage-dependent and non-linear ways, rather than a uniform offset along a fixed clock. Consequently, nominal time alone often fails to provide an interpretable and statistically valid quantification of condition-induced tempo changes. This motivates a shift from nominal-time staging alone toward embryo-resolved developmental-tempo inference.

Recent machine-learning approaches have made automated zebrafish analysis scalable. Earlier work used handcrafted features and conventional classifiers to assign embryos to discrete stages, demonstrating

* Corresponding author.
E-mail address: u8178636@anu.edu.au (L.-J.-Y. Deng).

https://doi.org/10.1016/j.ailsci.2026.100170
Received 7 February 2026; Received in revised form 9 April 2026; Accepted 9 April 2026
Available online 12 April 2026

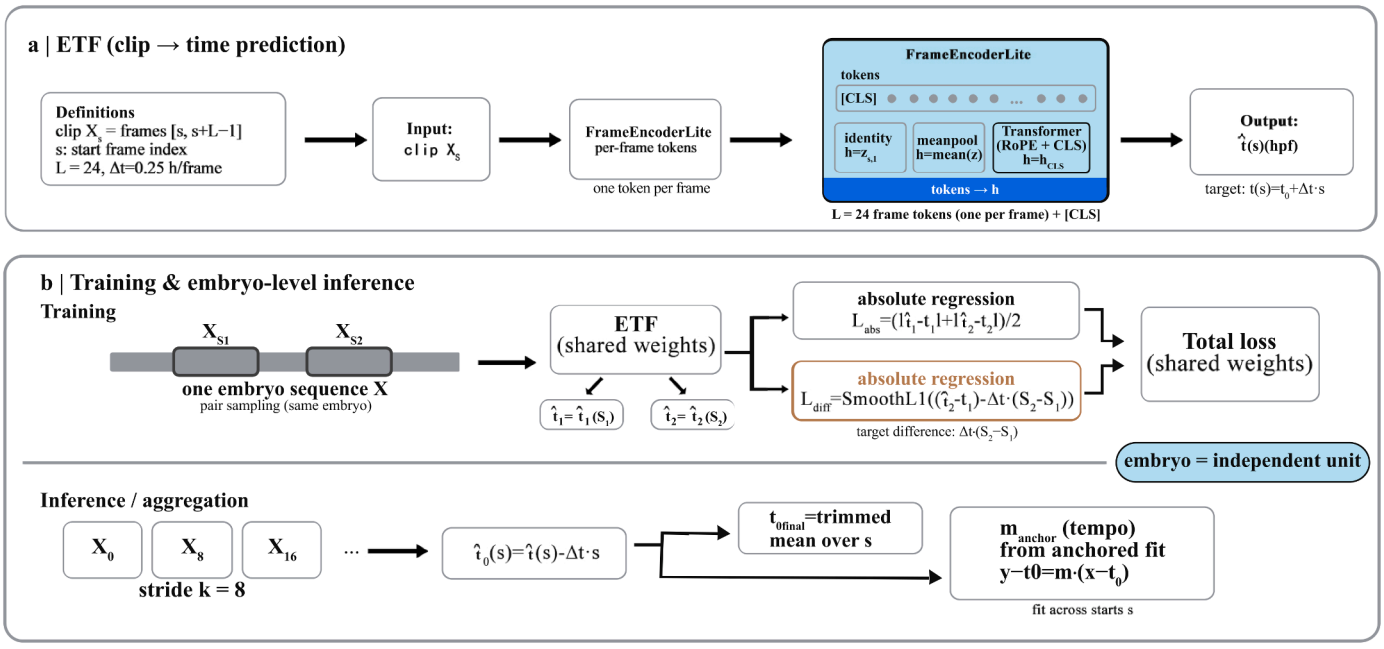


Fig. 1. Overview of EmbryoTempoFormer (ETF): clip-to-time prediction, within-embryo temporal-difference consistency training, and embryo-level tempo inference. ETF (the full model) predicts developmental time from clips extracted from zebrafish brightfield time-lapse sequences, with each clip containing $L = 24$ frames sampled at $\Delta t = 0.25\ h/frame$. During training, paired clips from the same embryo are optimized using an absolute regression objective together with a temporal-difference consistency regularizer, promoting coherent embryo-level trajectories across overlapping predictions. During inference, sliding-window predictions (stride = 8) are converted to $\hat{t}_0(s) = \hat{t}(s) - \Delta t{\cdot}s$ and robustly aggregated to obtain embryo-level readouts, including an anchored tempo slope m_anchor derived from an anchored fit at $T_0 = 4.5\ hpf$. Embryos, rather than frames or clips, are treated as the independent statistical units for downstream inferential analyses, thereby avoiding pseudo-replication [7].

feasibility but remaining sensitive to feature engineering choices imaging conditions [4]. More recently, deep-learning methods such as KimmelNet modeled staging as continuous regression by predicting developmental time directly from single 2D brightfield images, enabling automated staging and population-level delay detection [5]. Parallel work treated developmental time and tempo themselves as objects of inference, learning relationships among developmental states to reconstruct temporal mappings across conditions and revealing non-linear tempo behavior across developmental phases [6]. Together, these advances suggest that under systematic condition shifts (e.g., temperature changes), nominal hpf may be an unstable reference and an inadequate external "accuracy" target. These advances also highlight an unresolved opportunity for brightfield time-lapse data: temporal context within each embryo may contain information that is not fully captured by single-image staging alone. At the same time, the increasing availability of brightfield time-lapse microscopy introduces a practical inferential pitfall that can undermine statistical conclusions if left unaddressed. In practice, long sequences are often processed through densely sampled frames or overlapping sliding windows (clips) for training or inference. Predictions obtained from the same embryo are strongly correlated. Treating such correlated windows as independent samples for evaluation, hypothesis testing, or uncertainty estimation constitutes pseudo-replication and can lead to overconfident conclusions [7]. This issue becomes particularly pronounced under condition shifts, where nominal time is no longer a stable reference and window-level sample inflation can further distort downstream comparisons. A robust framework for time-lapse analysis should therefore (i) exploit temporal context, (ii) promote coherent within-embryo trajectories, and (iii) perform inference and uncertainty quantification at the embryo level.

Here we present the EmbryoTempoFormer (ETF), a clip-based framework for zebrafish brightfield time-lapse microscopy that predicts developmental progression from short temporal clips and promotes within-embryo temporal coherence during training through a temporal-difference consistency regularizer. In a downstream inference workflow, correlated clip-level predictions within each embryo are aggregated into embryo-level tempo and stability readouts suitable for cross-condition comparison, with embryos—rather than frames or clips—treated as the independent statistical units to avoid pseudo-replication [7]. This embryo-resolved framing is also compatible with the broader trend toward time-resolved phenotyping tasks such as developmental event detection, which aim to localize temporal phenotypes from microscopy sequences [8]. In the present study, we evaluate this framework in a matched two-condition temperature comparison (28.5 °C versus 25 °C), rather than across a broader perturbation panel. We therefore position the current manuscript as a temperature-shift proof-of-concept, while broader extension to additional temperatures, drug treatments, or environmental perturbations is reserved for future work with appropriately matched embryo-resolved time-lapse datasets. The main contributions of this work are:

(1) A clip-based CNN–Transformer framework that uses temporal context to predict developmental time from brightfield time-lapse microscopy.
(2) A within-embryo temporal-difference consistency regularizer that improves trajectory coherence across overlapping predictions.
(3) Embryo-level tempo readouts based on an anchored slope (m_anchor) and residual-based stability metrics for interpretable cross-condition comparisons.
(4) A reproducible end-to-end workflow and statistically rigorous cross-condition inference using embryo-bootstrap confidence intervals with embryos as the independent statistical units.

2. Related work

2.1. Automated zebrafish staging from images

Early automated staging approaches used handcrafted features and

Table 1
Embryo-level dataset splits and exclusions.

Condition (temperature)	Split type	Raw embryos (pre-filter)	Excluded embryos	Included embryos	Train embryos	Val embryos	Test embryos
25 °C (EXT25C)	test-only split	96	1	95	0	0	95
28.5 °C (ID28C5)	train/val/test split	192	51	141	98	21	22

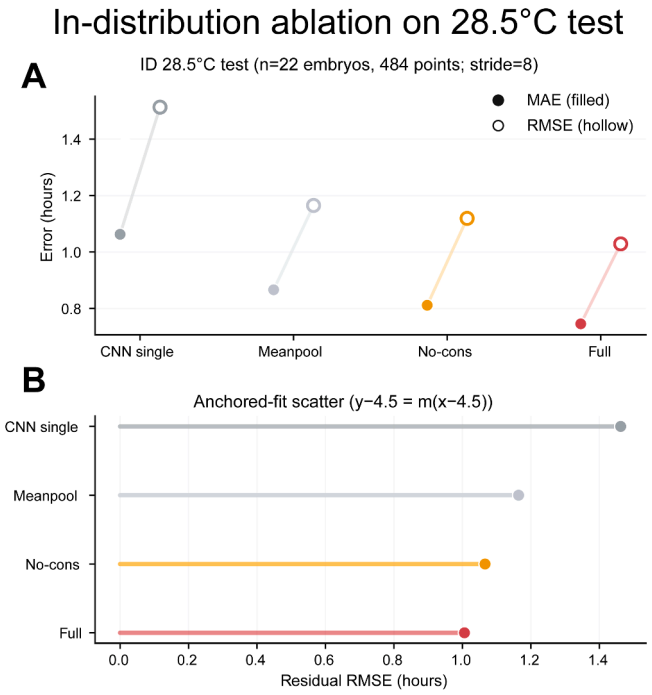


Fig. 2. Matched in-distribution ETF-family ablation on the 28.5 °C test set (ID28C5_TEST): descriptive clip-level error and embryo-level trajectory consistency.
(A) Clip-level MAE (filled) and RMSE (hollow) computed across all sliding-window clips (stride = 8; n = 484 correlated windows from n = 22 embryos). The four variants define a controlled temporal-modeling ladder: cnn_single is the matched image-only baseline, meanpool adds multi-frame context without explicit temporal-order modeling, nocons adds temporal Transformer aggregation, and full (ETF) further adds temporal-difference consistency regularization. Because windows within an embryo are correlated, these metrics are reported for descriptive comparison only.
(B) Anchored-fit residual RMSE (rmse_resid, in hours), summarizing how tightly sliding-window predictions align to a single anchored trajectory under the model y − T0 = m(x − T0) with T0 = 4.5 hpf (Section 3.9). Lower values indicate greater embryo-level trajectory consistency. Numerical values are reported in Table 2.

conventional machine-learning models to assign embryos to discrete stages, demonstrating feasibility but remaining sensitive to feature engineering choices and imaging variability [4]. More recently, deep learning has enabled continuous staging by regressing developmental time directly from single 2D brightfield images. KimmelNet is a representative example that achieves high agreement with expert staging and supports population-level detection of overall developmental delay or acceleration under standard conditions [5]. While effective, single-image regression does not explicitly exploit temporal context from time-lapse microscopy. It also does not by itself address the inferential issues that arise when dense sequences are analyzed through correlated frames or overlapping windows. This motivates clip-based modeling as a complementary direction, in which temporal context can be incorporated directly into developmental-time inference rather than added only at the level of downstream aggregation. A concise positioning of these method families relative to ETF is provided in Supplementary Table S1.

2.2. Developmental time and tempo under condition shifts

Beyond absolute time regression, recent work has treated developmental time and tempo themselves as objects of inference. By learning developmental-state similarity and reconstructing tempo mappings across conditions, such approaches capture non-linear characteristics of developmental progression and reveal tempo changes across developmental phases [6]. This perspective is particularly relevant under systematic condition shifts such as temperature changes, where nominal hpf may be an unstable alignment reference and "accuracy against a nominal clock" can be conceptually misleading. Our work is complementary to this line of research: we retain clip-to-time regression as a practical predictive primitive, while coupling it to embryo-level tempo readouts and statistically valid cross-condition comparison under correlated time-lapse sampling. This framing is especially useful when developmental delay is better interpreted as altered tempo rather than as a uniform temporal offset.

2.3. Deep learning for zebrafish phenotyping beyond staging

Deep learning has been applied to a broad range of zebrafish imaging tasks beyond staging, including phenotype recognition for high-throughput screening, deformation classification, anomaly detection of abnormal development, robustness-oriented multi-stage pipelines for brightfield larvae, integration with microfluidics for real-time and non-invasive embryo handling, and linking embryo phenotypes to mechanistic signaling pathways [9–14]. Reviews summarize this rapid expansion and also highlight persistent challenges in interpretability, robustness, and generalization. Time-resolved phenotyping has also expanded toward developmental event detection, where models identify the timing of specific developmental events from imaging sequences [8, 15]. Together directions underscore both the growing availability of dense imaging data and the need for principled temporal readouts that remain interpretable under condition shifts. ETF fits within this broader shift toward time-resolved phenotyping, but is specifically designed to quantify embryo-level developmental tempo rather than only detect isolated developmental events.

2.4. Time-lapse evaluation and pseudo-replication

Time-lapse sequences are frequently processed through densely sampled frames or overlapping sliding windows to support learning and inference over long recordings. However, windows from the same embryo are strongly correlated. Treating them as independent samples inflates effective sample size and constitutes pseudo-replication in statistical inference [7]. This pitfall is model-agnostic and becomes particularly consequential when downstream analyses report uncertainty or statistical significance using window-level samples rather than embryo-level units. ETF addresses this issue in two ways: by promoting within-embryo temporal coherence during training, and by aggregating correlated clip-level predictions into embryo-level readouts for inference and uncertainty quantification at the embryo level (Fig. 1).

3. Methods

3.1. Processing of OME-format microscopy data and NumPy preprocessing pipeline

Raw zebrafish brightfield time-lapse stacks were loaded from OME/

Table 2
In-distribution ETF-family ablation on ID28C5_TEST: descriptive clip-level error and anchored-fit consistency metrics. Table note: Data are derived from ID28C5_TEST (28.5 °C; $n = 22$ embryos). Sliding-window inference uses $L = 24$ frames, $\Delta t = 0.25$ h/frame, and stride $k = 8$, yielding $n = 484$ correlated windows. MAE, RMSE, and R^2 are computed over pooled window predictions and are reported for descriptive comparison only. Embryo-level summaries include the anchored tempo slope m_anchor, residual RMSE (rmse_resid), maximum absolute residual (max_abs_resid), and the 5th/95th percentiles of m_anchor across embryos. All values are rounded to three decimals.

Model	MAE (h)	RMSE (h)	R2	m_anchor (mean)	rmse_resid (h)	max_abs_resid (h)	m_P05	m_P95
full (ETF)	0.746	1.029	0.993	1.009	1.006	3.777	0.963	1.045
nocons	0.811	1.119	0.992	1.014	1.066	3.802	0.962	1.048
meanpool	0.867	1.165	0.992	1.002	1.165	5.328	0.95	1.041
cnn_single	1.063	1.513	0.986	0.984	1.462	7.116	0.933	1.022

TIFF files as multi-page sequences and represented as a NumPy array of shape [T, H, W]. For reproducibility, the corresponding processed arrays are distributed in the Zenodo bundle. Each embryo stack was converted to a fixed-shape uint8 array through percentile-based intensity normalization, spatial resizing, and temporal padding/trimming to a standardized sequence length. Briefly, intensities were clipped at percentile bounds (default 1–99) and linearly mapped to [0, 255], frames were resized to 384 × 384 pixels, and the temporal axis was standardized to 192 frames. One processed .npy file was stored per embryo, with embryo identifiers matched to file stems. During training and evaluation, processed arrays were accessed through NumPy memmap to enable efficient random access. Detailed preprocessing defaults and implementation notes are provided in Supplementary Table S1 and the released code.

3.2. Time base, clip definition, and sliding-window inference

For the released benchmark, we use the dataset-defined timing constants T0 = 4.5 hpf and Δt = 0.25 h/frame (15 min). A clip is defined as a contiguous window of length L extracted from a full embryo sequence:$Xs = \{frames[s], frames[s+1], \ldots, frames[s+L-1]\}$,

Where s is the clip start index and $L = 24$ by default. The nominal mapping from clip start index to nominal time is:

$$t(s) = T0 + \Delta t \cdot s.$$

During inference, we apply sliding windows with stride $k = 8$, yielding correlated clip-level predictions within each embryo. Embryos, rather than clips or windows, are treated as the independent statistical units for downstream inferential analyses thereby avoiding pseudo-replication [7].

3.3. Dataset splits, exclusions, and evaluation sets

All dataset splits were defined at the embryo level, with one embryo corresponding to one processed time-lapse stack. We evaluated ETF on two embryo-level test sets: ID28C5_TEST (28.5 °C), an in-distribution held-out test split with n = 22 embryos, and EXT25C_TEST (25 °C), a held-out temperature-shift test set with n = 95 embryos. For the 25 °C condition, the split file listed 96 processed embryos, of which 1 embryo was excluded according to a fixed filename rule (excluded file: FishDev_WT_25C_1_MMStack_G11-Site_0.ome), leaving 95 embryos in the final test set. For the 28.5 °C condition, the dataset initially comprised 192 raw embryo stacks from two 96-well plates. Following predefined filtering rules, 51 stacks were excluded because embryos were nonviable and/or outside the field of view, leaving 141 included embryos, which were split into train/validation/test sets of 98/21/22. These exclusion rules follow the original KimmelNet preprocessing pipeline and were additionally verified by manual review; exclusions were not performed post hoc to improve model performance [4]. The embryo-level dataset counts are summarized in Table 1.

3.4. Training and evaluation datasets

Training uses within-embryo paired sampling: each training sample consists of two clips (X_{s_1}, X_{s_2}) extracted from the same embryo sequence. This pairing construction supports the within-embryo temporal-difference consistency objective described in Section 3.7. To increase sampling diversity during training, clip start indices are jittered within a small range (default: start_jitter = 2). Evaluation uses a deterministic sliding-window protocol that enumerates all valid clip start indices for each embryo at a fixed stride (Section 3.2). Because sliding windows from the same embryo are strongly correlated, window-level metrics computed across all clips are reported for descriptive comparison only. All inferential analyses treat embryos, rather than clips or windows, as the independent statistical units, thereby avoiding pseudo-replication [7].

3.5. Data augmentation (training only)

Training-time augmentation was organized into four functional groups: spatial perturbations, photometric perturbations, acquisition-style perturbations, and temporal/sampling perturbations. Spatial transforms, including horizontal flips and mild affine transformations, were applied consistently across frames within a clip to preserve temporal coherence. Photometric perturbations included gamma, contrast, and brightness adjustments, together with low-frequency shading variation. Acquisition-style perturbations included additive noise and mild blur. Temporal/sampling perturbations included limited random frame drop during clip sampling; start-index jitter was applied separately at the sampling stage (Section 3.4). All augmentation magnitudes were intentionally kept mild to improve robustness to common brightfield-imaging variation while preserving morphology relevant for developmental staging. The same augmentation policy was used across all model variants (cnn_single, meanpool, nocons, and full). Augmented clips were clipped to [0, 1] before being passed to the model. Detailed augmentation defaults are summarized in Supplementary Table S2. The augmentation design was intended to simulate plausible real-world variation in embryo sampling and brightfield imaging conditions rather than to maximize the absolute score on a relatively clean benchmark split. Accordingly, some augmentation families may improve robustness to deployment-time variability while not necessarily yielding the best in-domain score on the present dataset. To assess this trade-off directly, we performed a matched five-setting remove-one-family sensitivity analysis for ETF-full, reported in Section 4.3. Here, no_temporal_sampling denotes removal of frame-drop perturbation and sampling-time start-index jitter; it does not disable the model-side temporal token dropout term.

3.6. Model architecture (EmbryoTempoFormer, ETF; full model)

EmbryoTempoFormer (ETF) predicts a scalar developmental time $\hat{t}(s)$ (hpf) from a grayscale clip X_s. For a batch of clips, the input tensor has shape $[B, L, 1, H, W]$, with default clip length $L = 24$.

Frame encoder (FrameEncoderLite). Each frame is encoded

Table 3
Four-level temporal-modeling ladder under the matched ETF protocol on ID28C5_TEST.

Incremental stage	Added capability	RMSE (h)	Relative RMSE improvement vs previous stage
cnn_single	Single-frame regression baseline	1.513	-
meanpool	24-frame context via direct averaging	1.165	23.0 %
nocons	Explicit inter-frame temporal modeling (Transformer)	1.119	3.9 %
full	Temporal-difference consistency regularization	1.029	8.1 %

independently into a token using a lightweight depthwise-separable CNN with GroupNorm and squeeze-and-excitation blocks [16–18]. The encoder outputs one token per frame, followed by global average pooling and a linear projection to token dimension D = 128. Thus, within-frame feature extraction is performed by the CNN frame encoder. For efficiency and memory control, frame encoding is performed in temporal chunks (frame_chunk) and can optionally use chunk-level frame-encoder checkpointing (ckpt_frame) under the selected memory profile. Temporal aggregation and model variants. Temporal aggregation is controlled by temporal_mode, yielding four model variants used throughout this study:

(1) cnn_single: image-only baseline; a single frame token is used for regression (no temporal context).
(2) meanpool: multi-frame baseline; frame tokens are averaged directly, without explicit temporal-order modeling.
(3) nocons: temporal Transformer aggregation without the temporal-difference consistency term (i.e., $\lambda_{\text{diff}} = 0$ in Section 3.7).
(4) full (ETF): temporal Transformer aggregation with the temporal-difference consistency term enabled ($\lambda_{\text{diff}} = 1$).

For Transformer-based variants (nocons and full), a learnable CLS token is prepended to the frame-token sequence and processed by a stack of Transformer encoder blocks with rotary positional embeddings (RoPE) applied to queries/keys [19,20]. The final CLS token is used for regression. In this architecture, the Transformer serves as an inter-frame temporal module operating across frame tokens. Training optionally applies temporal token dropout across frame tokens with rescaling to preserve expected magnitude.

Regression head: The regression head is Linear($D \to D$) → SiLU → Dropout → Linear($D \to 1$).

The number of trainable parameters is 0.276 M for cnn_single and meanpool and 0.804 M for nocons and full.

3.7. Training objective and optimization

Training samples are constructed by within-embryo paired sampling: two clips (X_{s_1}, X_{s_2}) are drawn from the same embryo, with clip start indices s_1 and s_2. The model outputs clip-level time predictions $\widehat{t}(s_1)$ and $\widehat{t}(s_2)$. The nominal target time for a start index s is $t(s) = T0 + \Delta t \cdot s$ (Section 3.2).

Absolute regression loss. Each clip was supervised using its nominal target using an elementwise regression loss $\ell(\cdot,\cdot)$ (L1 or SmoothL1/Huber) [21]:

$$L_{abs} = \frac{1}{2}\left(\ell(\widehat{t}(s_1), t(s_1)) + \ell(\widehat{t}(s_2), t(s_2))\right).$$

Within-embryo temporal-difference consistency. To encourage coherent within-embryo temporal progression across overlapping windows, we add a temporal-difference consistency term that constrains the predicted time difference between two clips from the same embryo to match the known sampling interval:

$$L_{diff} = \text{SmoothL1}\left(\left(\widehat{t}(s_2) - \widehat{t}(s_1)\right) - \Delta t \cdot (s_2 - s_1)\right).$$

Total loss with ramp-up. The overall training objective is:

$$L = \lambda_{abs} L_{abs} + \lambda_{diff}\, ramp(u)\, L_{diff},$$

where $ramp(u) \in [0,1]$ increases linearly from 0 to 1 over the first cons_ramp_ratio fraction of optimization steps (default 0.2). In the nocons ablation, $\lambda_{diff} = 0$; in the full (ETF) model, $\lambda_{diff} = 1$.

Optimization: All variants with AdamW and a cosine learning-rate schedule; the corresponding hyperparameter settings are summarized in Section 3.8 [22].

3.8. Training configuration

Released checkpoints store the full training configuration and are included in the Zenodo reproducibility bundle. Unless otherwise noted, all model variants were trained under the same configuration; the principal ablation-dependent setting was λdiff (0 for nocons, 1 for full; Section 3.7).

Training used clip length L = 24, image size 384 × 384, and expected sequence length T = 192. Models were trained for 300 epochs using within-embryo paired sampling, with start_jitter = 2 during training. Optimization used AdamW with learning rate 6×10^{-4} and weight decay 0.01, together with linear warm-up followed by cosine decay [22]. For Transformer-based variants, the default hyperparameters were model dimension D = 128, 4 attention heads, depth 4, and MLP ratio 2.0. Training used mixed precision and exponential moving average evaluation.

To improve memory efficiency at the reported input resolution and batch size, the released checkpoints used a memory-optimized configuration (mem_profile = "lowmem"), with chunked frame encoding and frame-encoder checkpointing enabled. Detailed training defaults, implementation-level configuration fields, and representative hardware/runtime requirements are provided in Supplementary Table S10 and the released code. A matched ETF-full augmentation-family sensitivity analysis was additionally performed under the same training framework and is reported in Section 4.3.

3.9. Embryo-level inference and tempo estimation

Sliding-window inference yields clip-level time predictions $\widehat{t}(s)$ (hpf) at multiple window start indices s within each embryo (Section 3.2). Because overlapping clips within an embryo are strongly correlated, these predictions are summarized into embryo-level readouts for downstream analysis.

Optional start-time offset diagnostic (descriptive only): Because fertilization and collection are not perfectly synchronized, the effective time zero can vary across embryos. As a descriptive diagnostic of this uncertainty, for each start index s we compute

$$\widehat{t}_0(s) = \widehat{t}(s) - \Delta t \cdot s,$$

where Δt is the sampling interval. The set $\{\widehat{t}_0(s)\}$ can be summarized within an embryo using a trimmed mean to obtain $t0_{\text{final}}$ (trim proportion = 0.2). This offset summary is reported for descriptive inspection only and is not used for inferential comparisons or hypothesis testing.

Primary tempo readout anchored slope m_{anchor}: Embryo-level developmental tempo is quantified using an anchored least-squares slope through the fixed anchor point $(T0, T0)$. For each index s, we define the nominal time coordinate

$$x(s) = T0 + \Delta t \cdot s,$$

and let

Table 4
Minimal augmentation-family sensitivity analysis for ETF-full under matched retraining settings.

Setting	ID28C5 MAE (h)	Change vs baseline_full	ID28C5 RMSE (h)	ID28C5 R2	EXT25C m_anchor	EXT25C rmse_resid (h)
baseline_full	0.818	0.0 %	1.126	0.992	0.710	1.353
no_spatial	1.324	+61.8 %	1.806	0.980	0.719	1.972
no_photometric	0.745	−9.0 %	1.075	0.993	0.790	1.487
no_acquisition	0.900	+10.0 %	1.198	0.991	0.709	1.516
no_temporal_sampling	0.749	−8.5 %	1.033	0.993	0.711	1.239

Table note: All runs used the same ETF-full architecture, split, optimizer, schedule, EMA configuration, and evaluation protocol; only the augmentation-family toggles differed. The reported change is relative to baseline_full on ID28C5_TEST. External 25 °C values are summarized using ETF's embryo-level tempo/stability readouts rather than interpreted as external-domain accuracy.

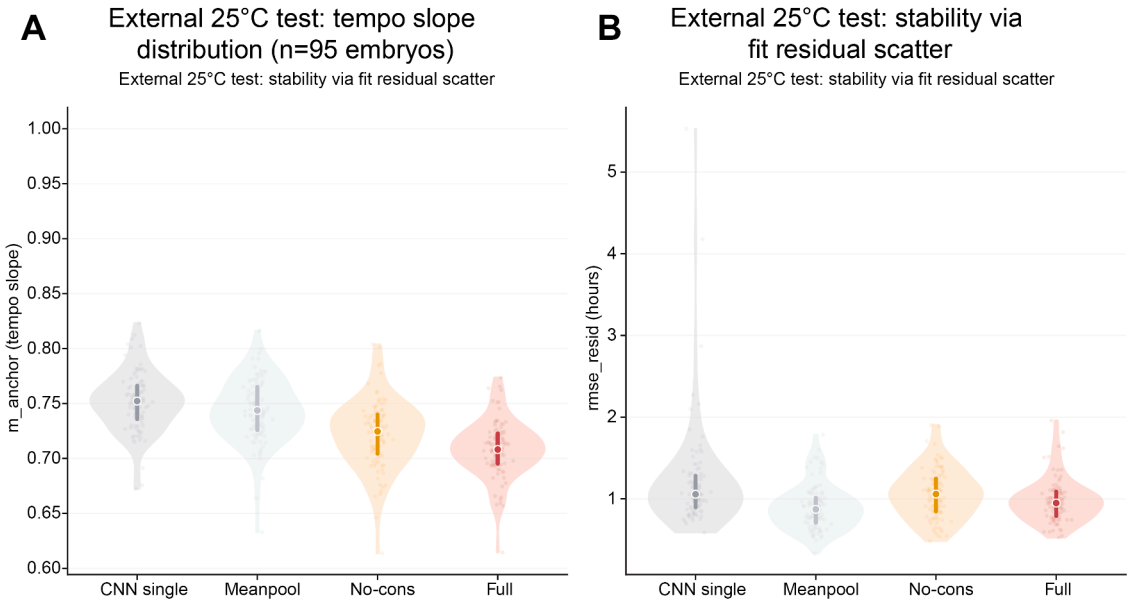


Fig. 3. Temperature-shift test set at 25 °C (EXT25C_TEST): embryo-level tempo distribution and residual-based stability across models.
(A) Distribution of embryo-level anchored tempo slopes m_{anchor} on EXT25C_TEST (25 °C; $n = 95$ embryos). Each point represents one embryo. Violins summarize the distribution across embryos; the white dot indicates the median and the thick vertical bar indicates the interquartile range (IQR, 25th–75th percentiles). The dashed line at $m = 1$ indicates nominal unit tempo.
(B) Distribution of per-embryo anchored-fit residual RMSE (rmse_resid, hours) on the same set, summarizing within-embryo trajectory self-consistency across sliding-window predictions (lower is better). Plot elements are as in (A): points are embryos; violins show density; white dot = median; thick bar = IQR. Tempo and residual metrics are computed per embryo from anchored fits at $T0 = 4.5$ hpf under the model $(y - T0) = m(x - T0)$ (Section 3.9). Numerical summaries are reported in Table 5. Under domain shift, meanpool shows a slightly lower median rmse_resid than the full model, consistent with the robustness of uniform temporal averaging to temperature-induced distribution shift.

$y(s) = \widehat{t}(s)$.

We then fit the anchored model

$$y - T0 = m_{\text{anchor}}(x - T0)$$

by least squares across starts s within the embryo, yielding the closed-form estimate

$$m_{\text{anchor}} = \frac{\sum_s (x(s) - T0)(y(s) - T0)}{\sum_s (x(s) - T0)^2}.$$

Residuals are defined as

$$r(s) = (y(s) - T0) - m_{\text{anchor}}(x(s) - T0),$$

from which we report embryo-level stability metrics: residual RMSE (rmse_resid, in hours) and maximum absolute residual (max_abs_resid, in hours). Under this definition, $m_{\text{anchor}} < 1$ indicates a slower tempo relative to the nominal axis, and $m_{\text{anchor}} > 1$ indicates faster tempo.

All embryo-level quantities (m_anchor, rmse_resid, and max_abs_resid) are computed separately for each embryo and used in downstream cross-condition analyses with embryos, rather than clips or windows, treated as the independent statistical units (Section 3.11), thereby avoiding pseudo-replication [7].

Anchor-sensitivity, stage-stratified error, and stage-dependent tempo summaries based on these embryo-level readouts are provided in Supplementary Tables S6–S8.

3.10. SmoothGrad saliency visualization (qualitative)

We use SmoothGrad saliency maps for qualitative inspection of which image regions contributed to clip-level time predictions [23]. For a clip X_s normalized to $[0, 1]$, saliency is defined as the absolute input gradient

$$G = \left| \frac{\partial \widehat{t}(s)}{\partial X_s} \right|.$$

SmoothGrad averages saliency over noisy perturbations of the input clip,

$$X_s^{(k)} = \text{clip}\left(X_s + \epsilon^{(k)}, 0, 1\right),\ \epsilon^{(k)} \sim N\left(0, \sigma^2\right),$$

yielding $G = \frac{1}{N} \sum_{k=1}^{N} G^{(k)}$. Unless otherwise noted, we use $N = 20$ and $\sigma = 0.01$.

Table 5
Temperature-shift test set at 25 °C (EXT25C_TEST): descriptive deviation from the nominal time axis and embryo-level tempo/stability summaries.

Model	MAE_vs_nominal (h)	RMSE_vs_nominal (h)	R2_vs_nominal	m_anchor (mean)	rmse_resid (h)	max_abs_resid (h)	m_P05	m_P95
full (ETF)	6.433	7.241	0.674	0.709	1.181	6.167	0.664	0.752
nocons	6.101	6.944	0.7	0.722	1.336	6.179	0.673	0.771
meanpool	5.568	6.409	0.745	0.743	1.169	5.444	0.703	0.786
cnn_single	5.253	6.286	0.755	0.751	1.52	17.435	0.716	0.801

Table note. Data are derived from EXT25C_TEST (25 °C; n = 95 embryos). Sliding-window inference used clip length L = 24, sampling interval Δt = 0.25 h/frame, stride k = 8, and anchor time T0 = 4.5 hpf. MAE_vs_nominal, RMSE_vs_nominal, and R^2_vs_nominal quantify deviation from the nominal mapping t(s) = T0 + Δt·s and are reported for descriptive comparison only. Primary embryo-level summaries include the anchored tempo slope m_anchor (reported as the mean across embryos, together with the 5th and 95th percentiles) and residual-based stability metrics (rmse_resid and max_abs_resid) derived from the same anchored fit. Sliding-window predictions are correlated within embryos; inferential analyses therefore treat embryos as the independent statistical units. Values are rounded to three decimals. *(Consistency note for readers: Fig. 3 displays medians and IQRs for visualization; this table reports means (and selected percentiles) for numeric summaries).*

Per-frame temporal importance is summarized as the spatial mean of G for each frame and min–max normalized within each clip for visualization. Heatmap rendering uses percentile stretching, blur, and alpha masking for readability only; these steps do not affect the underlying model predictions. Saliency results are reported as qualitative diagnostics and are not interpreted as causal explanations. Higher-resolution saliency overlays and zoomed-in views are provided in Supplementary Figures S3–S4.

3.11. Quantification and statistical analysis

Sliding-window clips within an embryo are strongly correlated. Therefore, in all inferential analyses, embryos, rather than clips or windows, are treated as the independent statistical units thereby avoiding pseudo-replication [7]. Window-level metrics computed over sliding windows are reported for descriptive comparison only and are not used for statistical inference.

Effect size (tempo): Temperature effects are quantified using embryo-level anchored tempo slopes m_{anchor} (Section 3.9). We define the tempo effect size as

$$\Delta m = \text{mean}\left(m_{\text{anchor}}^{25^\circ\text{C}}\right) - \text{mean}\left(m_{\text{anchor}}^{28.5^\circ\text{C}}\right).$$

Embryo-bootstrap confidence intervals: Confidence intervals are computed by embryo-level bootstrap resampling [24]. Embryos are resampled with replacement within each condition for $B = 5000$ replicates (seed = 0), and Δm is recomputed for each replicate. Two-sided 95 % confidence interval is obtained using the percentile method from the empirical bootstrap distribution of Δm.

Embryo-level sample-efficiency planning: We additionally evaluated embryo-budget requirements using the same embryo-level inferential framework, treating successful detection as exclusion of 0 from the bootstrap confidence interval for Δm. This analysis is intended as an inferential planning aid based on the observed dataset distribution, rather than as a prospective guarantee. Importantly, this analysis addresses embryo-budget requirements for inference and effect detection after model training. It is not a retraining-based learning-curve analysis of how many training embryos are required to fit the model, which would require repeated retraining across multiple training fractions and seeds under matched optimization settings. Detailed simulation settings are summarized in Supplementary Table S11.

4. Results

4.1. Overview: clip prediction and embryo-level tempo readouts

Fig. 1 summarizes the workflow from clip extraction to embryo-level tempo readouts. ETF predicts developmental time from short clips sampled from each brightfield time-lapse sequence. Unless otherwise noted, clips contain $L = 24$ frames sampled at $\Delta t = 0.25$ h/frame, and sliding-window inference uses stride $k = 8$, yielding clip-level predictions $\widehat{t}(s)$ at multiple starts indices s along each embryo sequence (Section 3.2).

Because sliding-window predictions within an embryo are strongly correlated, downstream analyses summarize clip-level outputs into embryo-level readouts (Section 3.9). Cross-condition comparisons focus on the anchored embryo-level tempo slope m_anchor together with residual-based stability metrics (rmse_resid and max_abs_resid) derived from anchored fits. In all inferential analyses, embryos, rather than clips or windows are treated as independent statistical units thereby avoid pseudo-replication in sliding-window time-lapse evaluation (Section 3.11) [7].

4.2. In-distribution ablation at 28.5 °C (ID28C5_TEST)

We first evaluated ETF (full) together with three matched ablations (cnn_single, meanpool, and nocons) on the 28.5 °C in-distribution test set (ID28C5_TEST; n = 22 embryos). Sliding-window inference used stride = 8, yielding 484 correlated clip-level predictions (Section 3.2). These four variants define a controlled temporal-modeling ladder under a shared protocol: cnn_single serves as the matched image-only baseline, meanpool adds multi-frame context without explicit temporal-order modeling, nocons adds temporal Transformer aggregation, and full further adds within-embryo temporal-difference consistency regularization. This design isolates the contributions of temporal context, adaptive temporal aggregation, and consistency regularization.

Under in-distribution conditions, clip-level MAE and RMSE computed across sliding-window clips provide a descriptive summary of time-prediction error. All clip-based variants outperform the image-only baseline, and ETF (full) achieves the lowest clip-level error among the tested models (Fig. 2A; Table 2). The same progression is summarized as a four-level temporal-modeling ladder in Table 3. Expanded metric summaries for the matched ETF-family benchmark are provided in Supplementary Table S4. Because windows from the same embryo are correlated, these pooled window-level metrics are reported for descriptive comparison only and are not used for inferential claims [7].

We therefore emphasize embryo-level trajectory consistency as a key property of time-lapse prediction. Using the anchored-fit framework defined in Section 3.9, ETF yields lower residual scatter (rmse_resid) and improved worst-case residual behavior (max_abs_resid) than simpler baselines (Fig. 2B; Table 2). In particular, the comparison between nocons and full isolates the contribution of the temporal-difference consistency regularizer: enabling consistency improves anchored-fit residual behavior while preserving the same embryo-level inferential unit.

4.3. Minimal augmentation-family sensitivity analysis

To quantify the contribution of the training-time augmentation pipeline described in Section 3.5, we performed a matched single-seed remove-one-family ablation for ETF-full under five settings: baseline_full, no_spatial, no_photometric, no_acquisition, and no_temporal_sampling. Accordingly, baseline_full in Table 4 should be interpreted as the matched retraining baseline for this sensitivity block,

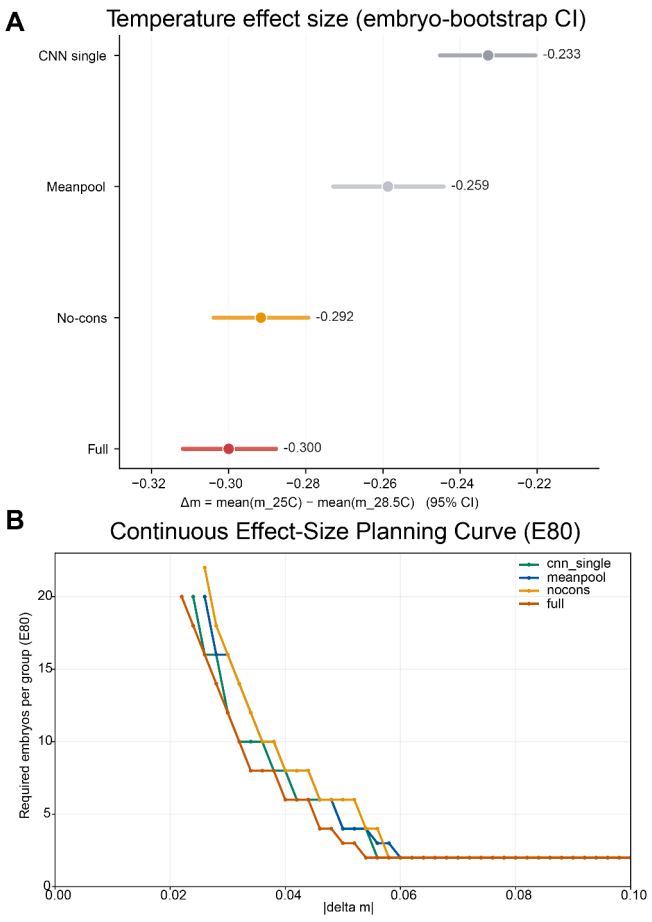


Fig. 4. Temperature effect size, embryo-bootstrap uncertainty, and embryo-budget planning.

(A) Temperature effect size Δm for each model, defined as the difference between mean embryo-level anchored tempo slopes at 25 °C and 28.5 °C. For each model, embryo-level tempo slopes m_anchor were computed separately for each embryo from anchored fits at T0 = 4.5 hpf under the model $y - T0 = m(x - T0)$ (Section 3.9). Points indicate observed Δm, and horizontal bars indicate embryo-bootstrap 95 % confidence intervals (B = 5000; seed = 0) obtained by resampling embryos with replacement within each condition (Section 3.11). Embryos, rather than clips or windows, were treated as the independent statistical units, thereby avoiding pseudo-replication [7]. Negative Δm indicates slower developmental tempo at 25 °C. Numerical values are reported in Table 6.

(B) Continuous embryo-budget planning curve (E80) across model regimes, showing the required embryos per condition for 80 % successful detection as a function of absolute effect size. All models show saturated detection for the observed large effect size, whereas meaningful separation appears in the subtle-to-moderate effect range. ETF (full) remains the most sample-efficient, or among the most sample-efficient, regimes in this range. A standalone high-resolution version of the all-model E80 planning curve is provided in Supplementary Figure S1, and the corresponding ETF-full E80/E90/E95 curves are provided in Supplementary Figure S2.

not as the released full-model result reported in Table 2. All runs used the same architecture, split, optimizer, learning-rate schedule, EMA configuration, memory profile, and evaluation protocol; only the augmentation-family toggles differed.

On the in-distribution test set (ID28C5_TEST), the clearest degradation was observed when removing the spatial family: MAE increased from 0.818 h in baseline_full to 1.324 h in no_spatial (+61.8 %). Removing the acquisition family had a smaller adverse effect (MAE 0.900 h; +10.0 %). By contrast, no_photometric and no_temporal_sampling yielded slightly lower in-distribution MAE than the matched retrained baseline (0.745 h and 0.749 h, respectively, versus 0.818 h).

These results indicate that augmentation-family effects are not uniformly positive or additive in this matched retraining block. The strongest clearly beneficial component is the spatial family, whereas the present photometric and temporal/sampling recipes may be somewhat over-regularizing for the current dataset. At the same time, lower in-distribution error did not automatically imply a preferable external tempo interpretation: for example, removing the photometric family shifted the external embryo-level tempo estimate toward weaker inferred slowdown (m_anchor 0.790 vs 0.710 for baseline_full on EXT25C_TEST). We therefore interpret this analysis as a family-sensitivity study rather than as a redefinition of the globally optimal training recipe. We did not extend this augmentation-family matrix to the Princeton cross-site experiments, because that would confound augmentation attribution with simultaneous site/domain shift; a broader cross-site augmentation-attribution study remains future work.

4.4. Temperature-shift test set at 25 °C: tempo distribution and stability (EXT25C_TEST)

We next analyze the 25 °C temperature-shift test set (EXT25C_TEST; $n = 95$ embryos; stride = 8; $n = 2090$ correlated windows). Under temperature shift, the nominal mapping from frame index to hpf is not a ground-truth developmental clock for 25 °C. Accordingly, point-level deviation from the nominal y = x mapping is interpreted descriptively rather than as accuracy against a ground-truth developmental time axis. We therefore focus on embryo-level tempo and residual-based stability readouts derived from anchored fits (Section 3.9).

Embryo-level tempo slopes at 25 °C show relatively tight distributions across embryos. The model-specific 5th–95th percentile ranges of m_{anchor} are: ETF 0.664–0.752, nocons 0.673–0.771, meanpool 0.703–0.786, and cnn_single 0.716–0.801 (Fig. 3A; Table 5). These embryo-level slopes provide an interpretable summary of developmental slowdown under temperature shift.

We further assess embryo-level trajectory stability using anchored-fit residual metrics (rmse_resid and max_abs_resid) (Fig. 3B; Table 5). These metrics complement tempo slopes by quantifying how coherently sliding-window predictions form a single anchored trajectory within each embryo and how sensitive each model is to outlier windows under domain shift. Notably, meanpool shows slightly lower rmse_resid than the full model (1.169 vs 1.181 h), consistent with a possible robustness advantage of uniform temporal averaging under this distribution shift, whereas cnn_single exhibits substantially heavier long-tail behavior (max_abs_resid = 17.435 h), indicating occasional highly inconsistent windows.

For completeness, Table 5 also reports descriptive deviation from the nominal 28.5 °C time axis. As expected under a temperature-induced tempo shift, models that infer stronger slowdown (lower m_anchor) also show larger deviation from the nominal axis. These window-level nominal-axis deviations are therefore descriptive only and are not interpreted as external-domain accuracy. The augmentation-family sensitivity analysis in Section 4.3 further shows that lower in-distribution clip-level error does not automatically imply a preferable external tempo interpretation. In particular, removing the photometric family reduced in-distribution error in the matched retraining block, but also shifted the external embryo-level tempo estimate toward weaker inferred slowdown. We therefore interpret the augmentation ablation as a sensitivity analysis of recipe components rather than as a new global ranking of universally optimal training settings.

4.5. Temperature effect size, embryo-bootstrap uncertainty, and sample-efficiency planning

To quantify the temperature-induced tempo shift with statistically valid uncertainty, we computed embryo-bootstrap confidence intervals treating embryos as independent statistical units (Section 3.11). We

Table 6
Temperature effect size Δm with embryo-bootstrap 95 % confidence intervals.

Model	Δm (25 °C − 28.5 °C)	95 % CI
cnn_single	−0.233	[−0.245, −0.221]
meanpool	−0.259	[−0.273, −0.244]
nocons	−0.292	[−0.304, −0.279]
full (ETF)	−0.300	[−0.312, −0.288]

Table note: Effect size definition: $\Delta m = \mathrm{mean}\left(m_{\mathrm{anchor}}^{25^{\circ}\mathrm{C}}\right) - \mathrm{mean}\left(m_{\mathrm{anchor}}^{28.5^{\circ}\mathrm{C}}\right)$, where m_{anchor} is the embryo-level tempo slope from an anchored fit at $T0 = 4.5$ hpf (Section 3.9). Confidence intervals are embryo-bootstrap 95 % CIs computed by resampling embryos with replacement within each condition (B = 5000; Section 3.11). Embryos, rather than sliding windows, were treated as the independent statistical units. Values are rounded to three decimals. Negative values indicate lower anchor-referenced tempo estimates at 25 °C.

define the temperature effect size as

$$\Delta m = \mathrm{mean}\left(m_{\mathrm{anchor}}^{25^{\circ}\mathrm{C}}\right) - \mathrm{mean}\left(m_{\mathrm{anchor}}^{28.5^{\circ}\mathrm{C}}\right),$$

where m_{anchor} is the embryo-level anchored tempo slope computed at $T0 = 4.5$ hpf (Section 3.9). Under this definition, negative Δm indicates slower developmental tempo at 25 °C.

Across all evaluated models, Δm is negative and the embryo-bootstrap 95 % confidence intervals lie entirely below zero (B = 5000; seed = 0), consistent with a robust slowdown under the anchored tempo definition (Fig. 4; Table 6). ETF shows the largest effect magnitude ($\Delta m = -0.300$, CI95 $[-0.312, -0.288]$; Table 6).

We further evaluated embryo-budget requirements using the same embryo-level inferential framework. For the observed large effect size (approximately $|\Delta \mathrm{m}| \approx 0.30$), detection was effectively saturated across all models at very small embryo budgets. The planning analysis was therefore most informative in the subtle-to-moderate effect range, where differences between model regimes became visible. In this regime, ETF (full) remained the most sample-efficient, or among the most sample-efficient, variants under increasingly stringent detection targets. The all-model continuous E80 planning curve is shown in Fig. 4B, and the corresponding ETF-full E80/E90/E95 curves are provided in Supplementary Figure S2. Representative embryo-budget planning points are summarized in Supplementary Table S5.

4.6. Cross-site validation on Princeton data

This cross-site analysis complements, rather than duplicates, the held-out 25 °C temperature-shift test set because the Princeton data introduce site-level acquisition variation rather than a controlled temperature perturbation within the same benchmark context.

To assess whether the matched ETF framework remained informative beyond the released benchmark site, we additionally evaluated a genuine cross-site dataset from Princeton (S-BIAD840). This experiment was designed to complement, rather than duplicate, the held-out 25 °C temperature-shift test set, because the Princeton data introduce acquisition-domain and site-level variation rather than a controlled temperature perturbation within the same benchmark context.

Under strict zero-shot transfer, all frozen ETF-family variants exhibited a clear calibration shift on Princeton data, consistent with substantial domain mismatch in image scale, background appearance, illumination, and overall visual texture. For readability, the main text focuses on the matched image-only baseline (cnn_single) and the main clip-based model (full). In this zero-shot setting, ETF (full) outperformed cnn_single on both Princeton test sets, reducing RMSE from 14.339 h to 9.862 h on SBIAD840_28C5_TEST and from 14.116 h to 10.643 h on SBIAD840_25C_TEST (Table 7). However, all frozen models remained clearly miscalibrated relative to in-site performance, indicating that cross-site zero-shot transfer is not yet solved in this setting.

We therefore tested a constrained low-shot adaptation strategy on two representative models. Using only 12 Princeton 28.5 °C embryos for fine-tuning, performance improved substantially for both models. On SBIAD840_28C5_TEST, RMSE decreased from 14.339 h to 5.270 h for cnn_single and from 9.862 h to 3.556 h for full. On SBIAD840_25C_TEST, adaptation reduced RMSE from 14.116 h to 5.017 h for cnn_single and from 10.643 h to 5.218 h for full. Although adapted cnn_single achieved slightly lower RMSE than adapted full on Princeton 25 °C, adapted full retained lower residual dispersion and higher through-origin line-fit R^2, while showing the strongest overall recovery on Princeton 28.5 °C. Expanded Princeton results and the adaptation protocol are provided in Supplementary Tables S9 and S12, respectively.

Overall, the Princeton analysis supports the utility of ETF beyond the original benchmark site while also defining a clear claim boundary: true zero-shot transfer across microscopy sites remains limited mainly by calibration shift, whereas modest site-specific fine-tuning can recover much of the lost performance. We therefore interpret this experiment as an initial cross-site validation rather than a claim of broad site-invariant generalization.

4.7. Temporal evidence within clips (interpretability)

To provide a qualitative interpretability check and to examine whether temporal aggregation differs from uniform averaging, we visualized within-clip temporal importance using SmoothGrad (Fig. 5; Section 3.10). We compute pixel-level saliency as the absolute input gradient $|\partial \hat{t}(s)/\partial X_s|$ and average it over noisy input perturbations (SmoothGrad). For each frame, we summarize saliency by the spatial mean of the absolute gradient and apply within-clip min–max normalization for visualization, yielding a temporal importance profile (Fig. 5A). Under this normalization, the curve reflects relative within-clip sensitivity across frames rather than absolute saliency magnitude.

Across five representative windows spanning early to late development (start = 0/42/84/126/168), temporal importance exhibits pronounced non-uniform peaks, and peak locations varied with developmental phase (Fig. 5A). The corresponding spatial overlays (Fig. 5B–F) suggested a phase-dependent shift in sensitivity patterns: early windows emphasized yolk-related morphology and the embryo–yolk boundary, early-to-mid windows increasingly highlighted axial and caudal structures alongside persistent yolk-related cues, and later windows showed more localized sensitivity around head and eye regions together with residual yolk-related signal. This qualitative progression is consistent with the increasing availability of localized morphological landmarks over development.

These observations provide a qualitative rationale for moving beyond mean pooling: peaked temporal importance profiles indicate that within-clip evidence is not uniformly distributed across frames, suggesting that some frames may carry more informative cues for time estimation. Mean pooling enforces equal weighting across frames and may dilute such informative signals, whereas Transformer-based aggregation can represent non-uniform evidence usage. We emphasize that SmoothGrad saliency reflects local sensitivity rather than causal attribution; we therefore use it as a qualitative interpretability check rather than a mechanistic explanation [23].

5. Discussion

A central challenge in zebrafish brightfield time-lapse analysis is that sliding-window inference produces many highly correlated predictions from the same embryo. Treating these window-level outputs as independent samples inflates the effective sample size and constitutes pseudo-replication, leading to overconfident conclusions [7]. In this work, EmbryoTempoFormer (ETF) produces clip-level developmental time predictions $\hat{t}(s)$ for window start index s, while the downstream inference and statistical workflow aggregates these temporally correlated window predictions into embryo-level tempo readouts and

Table 7
Princeton cross-site zero-shot and low-shot adaptation summary for the matched image-only baseline and ETF.

External dataset	Model / setting	RMSE (h)	Through-origin slope (m_origin)	Residual mean ± SD (h)	Through-origin line-fit R^2
SBIAD840_28C5_TEST	cnn_single zero-shot	14.339	1.288	5.07 ± 11.31	−3.611
SBIAD840_28C5_TEST	full zero-shot	9.862	1.203	2.78 ± 7.99	0.179
SBIAD840_28C5_TEST	cnn_single + ft12 + frame_tail1	5.270	1.003	0.98 ± 5.18	0.732
SBIAD840_28C5_TEST	full + ft12 + temporal_last1	3.556	0.948	0.47 ± 3.28	0.889
SBIAD840_25C_TEST	cnn_single zero-shot	14.116	1.260	5.09 ± 11.44	−3.783
SBIAD840_25C_TEST	full zero-shot	10.643	1.151	3.50 ± 9.31	−0.531
SBIAD840_25C_TEST	cnn_single + ft12 + frame_tail1	5.017	0.918	0.84 ± 4.50	0.759
SBIAD840_25C_TEST	full + ft12 + temporal_last1	5.218	0.862	0.84 ± 3.81	0.785

Note: For cross-paper comparability to the Princeton/KimmelNet reporting convention, this table uses the auxiliary through-origin view defined by y = mx. This is distinct from ETF's native anchored summary, which uses the anchor-referenced fit described in Section 3.9. Expanded Princeton zero-shot and adaptation summaries for the matched image-only baseline and ETF are provided in Supplementary Table S9, and the adaptation protocol is summarized in Supplementary Table S12.

quantifies uncertainty using embryos—rather than windows—as the independent statistical units. To avoid mixing statistical levels, we explicitly separate (i) window-level error metrics used for descriptive model comparison from (ii) embryo-level tempo and stability readouts used for interpretation and inference.

The ablation results disentangle three ingredients for time-lapse staging—temporal context, adaptive temporal evidence fusion, and within-embryo temporal consistency—through a staged comparison (cnn_single → meanpool → nocons → full; Table 2). First, adding short temporal context (cnn_single → meanpool) improves descriptive clip-level accuracy on ID28C5_TEST (MAE 1.063 → 0.867 h; RMSE 1.513 → 1.165 h; Table 2), supporting the premise that short temporal windows contain information beyond single-frame morphology. Second, beyond average error, anchored-fit residual behavior summarizes whether overlapping windows align with a coherent within-embryo trajectory—an important property when dense sliding-window predictions are interpreted as a single embryo's developmental course under correlated sampling [7]. Relative to meanpool, adaptive temporal aggregation (meanpool → nocons) improves both error and residual behavior (MAE 0.867 → 0.811 h; rmse_resid 1.165 → 1.066 h; max_abs_resid 5.328 → 3.802 h; Table 2), consistent with non-uniform evidence fusion reducing inconsistent or outlier windows. Third, comparing nocons and full isolates the contribution of the temporal-difference consistency regularizer: enabling consistency improves both average error and anchored-fit residual scatter (MAE 0.811 → 0.746 h; rmse_resid 1.066 → 1.006 h; Table 2), while leaving worst-case residuals nearly unchanged (max_abs_resid 3.802 → 3.777 h). Together, these results motivate reporting not only pointwise error but also trajectory-level self-consistency when interpreting dense time-lapse predictions produced by overlapping sliding windows.

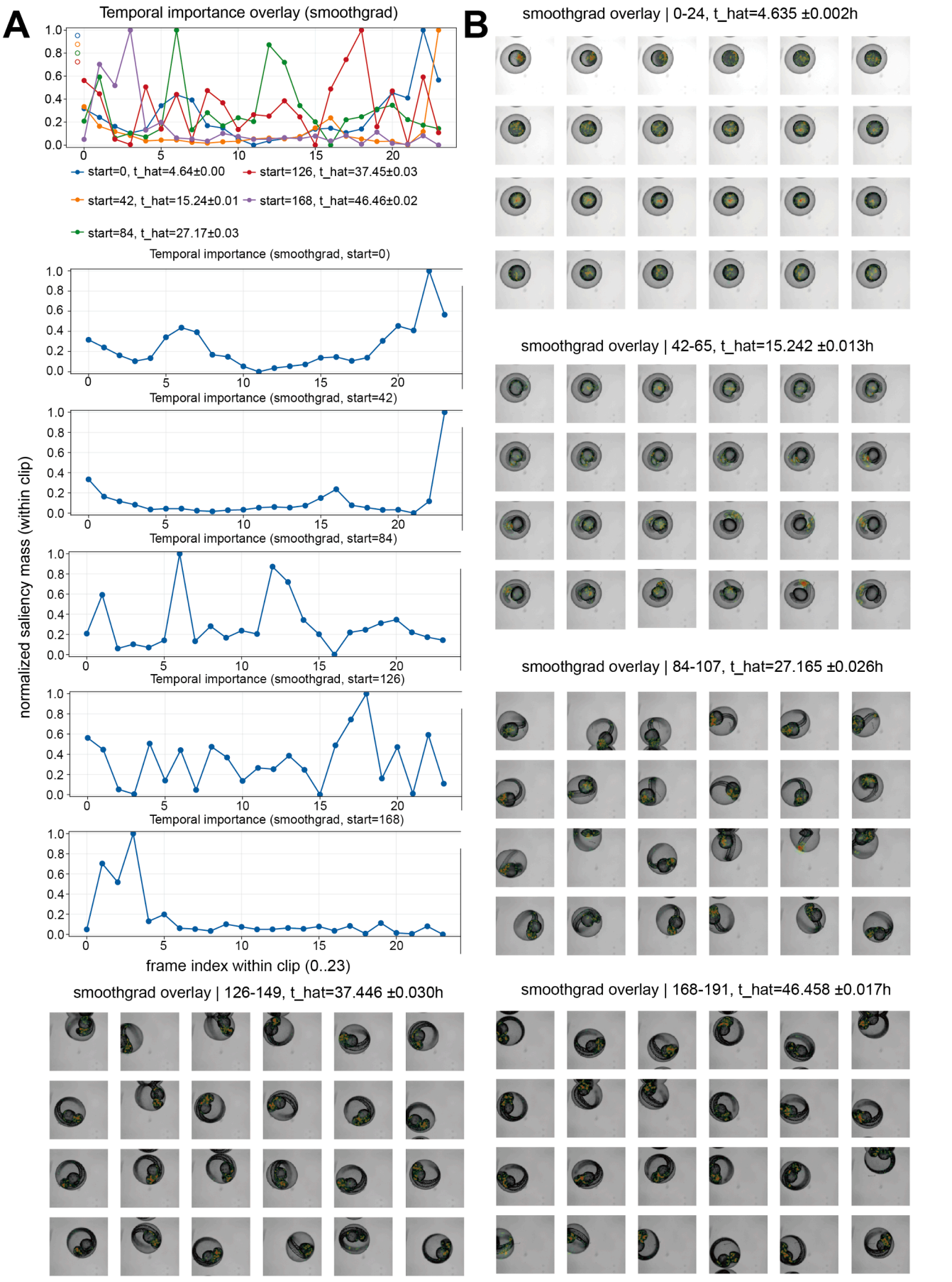
The augmentation-family sensitivity analysis further clarifies how the training recipe should be interpreted. In the present matched retraining block, spatial perturbations were clearly beneficial, whereas the removal effects of other families were smaller, neutral, or even favorable under some readouts. This behavior is consistent with augmentation interactions being non-additive: a perturbation family that is well motivated as a robustness prior need not maximize the absolute score on a relatively clean benchmark split. We therefore interpret the current augmentation analysis as a directional family-sensitivity study, not as a claim that every augmentation family contributes positively or that the present default recipe is globally optimal in all settings.

Embryo-level tempo readouts become more informative under domain shift, where a nominal clock is not an appropriate external accuracy target. Under temperature change, the nominal mapping from frame index to hpf no longer represents a ground-truth developmental clock for 25 °C. We therefore focus on embryo-level anchored tempo slopes (reported as m_anchor in tables) and residual-based stability metrics as the primary external readouts (Fig. 3; Table 5). Under this framing, deviation from the nominal mapping reflects accumulated deviation from the nominal axis rather than external-domain staging accuracy. This perspective is consistent with recent work emphasizing developmental time and tempo as objects of inference under condition shifts [6]. On EXT25C_TEST, m_anchor distributions are tight across embryos (ETF P05–P95: 0.664–0.752; Table 5), enabling interpretable embryo-resolved summaries of developmental slowdown under temperature shift.

Crucially, uncertainty for the temperature effect can be stated at the correct inferential level by treating embryos as independent units. Across all evaluated models, embryo-bootstrap 95 % confidence intervals for the effect size Δm = mean(m_anchor at 25 °C) − mean (m_anchor at 28.5 °C) lie entirely below zero (Table 6), consistent with a robust slowdown under the anchored tempo definition. ETF yields the largest effect magnitude (Δm = −0.300, CI [−0.312, −0.288]; Table 6). Embryo-level bootstrap resampling provides a principled uncertainty estimate when embryos are the independent units and avoids pseudo-replication that would arise from resampling correlated windows [7, 24]. At the same time, external vs_nominal metrics should not be interpreted as accuracy: models that infer a stronger slowdown (lower m_anchor) naturally tend to show larger deviation from the 28.5 °C nominal axis (e.g., ETF: m_anchor = 0.709, MAE_vs_nominal = 6.433 h; cnn_single: m_anchor = 0.751, MAE_vs_nominal = 5.253 h; Table 5). Residual-based stability provides complementary information about robustness under shift: meanpool shows slightly lower rmse_resid than ETF at 25 °C (1.169 vs 1.181 h; Table 5), suggesting a mild robustness advantage of uniform averaging under distribution shift, whereas the much larger max_abs_resid for cnn_single (17.435 h; Table 5) indicates occasional highly inconsistent windows when temporal modeling is insufficient. For presentation, Fig. 3 visualizes embryo-level distributions using medians and IQRs, whereas Table 5 reports numeric summaries computed from embryo-level aggregation; these views are complementary.

Together, these results motivate reporting tempo and stability at the embryo level rather than relying on window-level deviation from a nominal clock under condition shifts. Qualitative interpretability further contextualizes why uniform temporal averaging can be limiting. SmoothGrad analyses show non-uniform, phase-dependent temporal importance within 24-frame clips (Fig. 5A), with spatial sensitivity patterns shifting across development (Fig. 5B–F; Supplementary Figures S3 and S4). While saliency reflects local sensitivity rather than causal attribution, these structured patterns are compatible with the intuition that informative morphological landmarks emerge and change over time, supporting aggregation mechanisms that can represent non-uniform evidence usage [23].

Overall, this work complements prior single-image staging approaches such as KimmelNet by leveraging short temporal context and, critically, by centering embryo-resolved inference and embryo-level statistics to avoid pseudo-replication [5,7]. The approach integrates (i) clip-based modeling to provide local temporal context, (ii) adaptive temporal evidence fusion, (iii) within-embryo temporal-difference consistency to promote coherent trajectories, and (iv) embryo-level tempo readouts with embryo-bootstrap uncertainty to support rigorous, interpretable comparisons across conditions. Limitations are modest but



(caption on next page)

Fig. 5. SmoothGrad suggests non-uniform, phase-dependent evidence usage within 24-frame clips.
(A) Within-clip temporal importance profiles for five clip starts (start = 0/42/84/126/168). For each frame, importance is computed as the spatial mean of SmoothGrad saliency $G = |\partial\widehat{t}(s)/\partial X_s|$, where SmoothGrad averages saliency over $N = 20$ noisy perturbations with $\sigma = 0.01$ on inputs normalized to $[0,1]$. Curves are min–max normalized within each clip for visualization, reflecting relative within-clip sensitivity. (B–F) SmoothGrad heatmap overlays for the corresponding 24-frame clips at the five starts (one panel per start; 24 frames per panel). Heatmaps are visualized with percentile stretching, blur, and alpha masking for readability; these rendering steps affect visualization only. Saliency reflects local sensitivity rather than causal attribution [23].

important. External validation in this revision includes both the held-out temperature-shift test set and an initial real cross-site evaluation on Princeton data. Unlike the held-out 25 °C temperature-shift analysis, the Princeton experiment probes true acquisition-domain shift across sites. However, broader generalization across imaging setups, laboratories, and acquisition protocols still remains to be tested. Biologically, the current results directly establish only a matched two-condition temperature comparison (28.5 °C versus 25 °C). Extending this scope to additional temperatures or non-temperature perturbations would require matched embryo-resolved time-lapse datasets with aligned sampling cadence, annotation semantics, and experimental metadata. In addition, the anchored tempo definition depends on the chosen reference T0, although the qualitative conclusions were stable across the anchor-sensitivity analysis reported in Supplementary Table S6; multi-anchor or nonlinear tempo formulations could be explored to assess sensitivity while retaining embryo-level inference. Finally, SmoothGrad is qualitative and method-dependent; it is included as a compatibility check for non-uniform temporal evidence usage rather than a mechanistic explanation [23].

6. Conclusions

EmbryoTempoFormer (ETF) is a clip-based CNN–Transformer model for developmental-time inference from zebrafish brightfield time-lapse microscopy. In this study, we paired ETF with an embryo-resolved inference and statistical workflow that aggregates temporally correlated sliding-window predictions into interpretable embryo-level tempo and stability readouts, and quantifies cross-condition effects using embryo-bootstrap confidence intervals with embryos treated as the independent units. Across the matched temporal-modeling ablation ladder, adding temporal context, explicit inter-frame temporal modeling, and temporal-difference consistency progressively improved descriptive prediction error and embryo-level trajectory stability. Under a matched two-condition temperature-shift setting, embryo-level tempo summaries and embryo-bootstrap uncertainty estimates yielded interpretable condition-level comparisons, while the added Princeton analysis provided an initial cross-site generalization test. Together, these results support ETF as a practical framework for embryo-resolved developmental-tempo analysis in brightfield time-lapse microscopy, while broader extension to additional temperatures or treatment conditions remains contingent on matched embryo-resolved time-lapse datasets.

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact: Li-Jia-Yu Deng (u8178636@anu.edu.au).

Experimental model and subject details

We analyzed publicly available zebrafish brightfield time-lapse microscopy data from the BioImage Archive, comprising the released benchmark dataset S-BIAD531 and an additional cross-site dataset S-BIAD840. Within S-BIAD531, embryos imaged at 28.5 °C were used for training and in-distribution evaluation, whereas embryos imaged at 25 °C were reserved as a held-out temperature-shift test set. The S-BIAD840 dataset was used separately for real cross-site validation. Embryo-level dataset splits and exclusions are summarized in Table 1. No new animal experiments were performed in this work.

Data and code availability

Code repository: https://github.com/LijiayuDeng/s-biad531-embryo-tempoformer.

Zenodo reproducibility bundle (processed arrays, model checkpoints, dataset splits, and checksums): https://doi.org/10.5281/zenodo.18318139.

Raw benchmark data: BioStudies accession S-BIAD531: https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD531.

Additional cross-site data: BioStudies accession S-BIAD840: https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD840?query=S-BIAD840.

Additional cross-site reproducibility record: Zenodo record 18,979,476: https://zenodo.org/records/18979476.

All quantitative results and figures can be reproduced using the provided scripts.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used the OpenAI model gpt-5.2-2025-12-11 to assist with language editing, manuscript organization/formatting, and code suggestions/refactoring for analysis and visualization scripts. After using the tool, the authors reviewed, verified, and edited the material as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Li-Jia-Yu Deng: Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Project administration, Writing – original draft, Writing – review & editing. **Pei-Ran Lin:** Conceptualization, Resources, Writing – review & editing. **Luo-Tong Xie:** Data curation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2026.100170.

Data availability

Data will be made available on request.

References

[1] Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Dev Dyn 1995;203(3):253–310. https://doi.org/10.1002/aja.1002030302.

[2] Parichy DM, Elizondo MR, Mills MG, Gordon TN, Engeszer RE. Normal table of postembryonic zebrafish development: staging by externally visible anatomy of the living fish. Dev Dyn 2009;238(12):2975–3015. https://doi.org/10.1002/dvdy.22113.

[3] Singleman C, Holtzman NG. Growth and maturation in the zebrafish, Danio rerio: a staging tool for teaching and research. Zebrafish 2014;11(4):396–406. https://doi.org/10.1089/zeb.2014.0976.
[4] Jones RA, Renshaw MJ, Barry DJ, Smith JC. Automated staging of zebrafish embryos using machine learning. Wellcome Open Res 2022;7:275. https://doi.org/10.12688/wellcomeopenres.18313.3.
[5] Jones RA, Renshaw MJ, Barry DJ. Automated staging of zebrafish embryos with deep learning. Life Sci Alliance 2024;7(1):e202302351. https://doi.org/10.26508/lsa.202302351.
[6] Toulany N, Morales-Navarrete H, Čapek D, Grathwohl J, Ünalan M, Müller P. Uncovering developmental time and tempo using deep learning. Nat Methods 2023;20(12):2000–10. https://doi.org/10.1038/s41592-023-02083-8.
[7] Hurlbert SH. Pseudoreplication and the design of ecological field experiments. Ecol Monogr 1984;54(2):187–211. https://doi.org/10.2307/1942661.
[8] Ibbini Z, Truebano M, Spicer JI, McCoy JCS, Tills O. Dev-ResNet: automated developmental event detection using deep learning. J Exp Biol 2024;227(10):jeb247046. https://doi.org/10.1242/jeb.247046.
[9] Wang B, Sun Q, Liu Y, Zhang J, Li G, Wu S, Zheng H, Ye J, Zhou M, Zheng H, Yu Y, Zhong Y, Wu Y, Huang D, Weng Z. Intelligent larval zebrafish phenotype recognition via attention mechanism for high-throughput screening. Comput Biol Med 2025;188:109892. https://doi.org/10.1016/j.compbiomed.2025.109892.
[10] Ishaq O, Sadanandan SK, Wählby C. Deep fish: deep learning–based classification of zebrafish deformation. SLAS Discov 2017;22(1):102–7. https://doi.org/10.1177/1087057116667894.
[11] Shang S, Lin S, Cong F. Zebrafish larvae phenotype classification from bright-field microscopic images using a two-tier deep-learning pipeline. Appl Sci 2020;10(4):1247. https://doi.org/10.3390/app10041247.
[12] Sivaprasad S, Wang H-P, Jäckel A-L, Baumann J, Baumann C, Herrmann J, Fritz M. Automated detection of abnormalities in zebrafish development. Medical image computing and computer assisted intervention – MICCAI 2025. lecture notes in computer science. 2025. https://doi.org/10.1007/978-3-032-04981-0_6.
[13] Diouf A, Sadak F, Bereziat L, Gerena E, Fage F, Mannioui A, Zizioli D, Fassi I, Boudaoud M, Legnani G, Haliyo S. Combining deep learning and microfluidics for fast and noninvasive sorting of zebrafish embryo. Sci Rep 2025;15(1):37778. https://doi.org/10.1038/s41598-025-17946-7.
[14] Čapek D, Safroshkin M, Morales-Navarrete H, Toulany N, Arutyunov G, Kurzbach A, Bihler J, Hagauer J, Kick S, Jones F, Jordan B, Müller P. EmbryoNet: using deep learning to link embryonic phenotypes to signaling pathways. Nat Methods 2023;20(6):815–23. https://doi.org/10.1038/s41592-023-01873-4.
[15] Mapstone C, Plusa B. Machine learning approaches for image classification in developmental biology and clinical embryology. Development 2025;152(4):dev202066. https://doi.org/10.1242/dev.202066.
[16] Chollet F. Xception: deep learning with depthwise separable convolutions. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR); 2017. https://doi.org/10.48550/arXiv.1610.02357. arXiv:1610.02357.
[17] Wu Y, He K. Group normalization. Computer vision – ECCV. 2018. https://doi.org/10.1007/978-3-030-01261-8_1. 2018.
[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2018. https://doi.org/10.48550/arXiv.1709.01507. arXiv:1709.01507.
[19] Su J, Ahmed MHM, Lu Y, Pan S, Bo W, Liu Y. RoFormer: enhanced transformer with rotary position embedding. Neurocomputing 2024;568:127063. https://doi.org/10.1016/j.neucom.2023.127063.
[20] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. 2017, https://doi.org/10.48550/arXiv.1706.03762, arXiv:1706.03762.
[21] Huber PJ. Robust estimation of a location parameter. In: Kotz S, Johnson NL, editors. Breakthroughs in statistics: methodology and distribution. New York, NY: Springer; 1992. p. 492–518. https://doi.org/10.1007/978-1-4612-4380-9.
[22] Loshchilov I., Hutter F. Decoupled weight decay regularization. 2017, https://doi.org/10.48550/arXiv.1711.05101, arXiv:1711.05101.
[23] Smilkov D., Thorat N., Kim B., Viégas F.B., Wattenberg M. SmoothGrad: removing noise by adding noise. 2017, https://doi.org/10.48550/arXiv.1706.03825, arXiv:1706.03825.
[24] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall/CRC; 1993. https://doi.org/10.1201/9780429246593.