# Video Diffusion based Digital Twin for Large Format Additive Manufacturing

Lu Liu, Haoyang Xie, Dylan Hoskins, Kyle Rowe, Feng Ju

***Abstract*— Large Format Additive Manufacturing (LFAM) enables the fabrication of large, complex structures but presents challenges in thermal management, particularly in determining the optimal layer time to ensure interlayer bonding and structural integrity. Digital Twin (DT) technology has emerged as a key solution for predicting temperature distributions and optimizing process parameters. However, existing Physics-Based and Data-Driven DT models provide static, one-time predictions, lacking the adaptability to dynamically update thermal profile predictions based on real-time parameter adjustments. To address this limitation, we propose an adaptive Digital Twin framework based on the Video Diffusion Transformer (VDT). Unlike traditional DT models, our approach leverages Generative AI to dynamically simulate future temperature distributions when layer time or other printing parameters change. This method ensures that adjustments in printing strategy are immediately reflected in updated temperature predictions, leading to enhanced efficiency, improved print quality, and greater adaptability in LFAM workflows. Experimental results demonstrate that our approach is highly effective, generating realistic future frames that accurately reflect the temperature distribution. This work represents a significant step forward in Digital Twin technology, highlighting the potential of Generative AI in manufacturing.**

## I. INTRODUCTION

Large Format Additive Manufacturing (LFAM) enables the fabrication of large, complex structures but introduces challenges in thermal management. A key factor affecting print quality is layer time, which determines interlayer bonding strength and overall structural integrity [1]. If the layer time is too short, excessive heat retention may cause deformation and poor adhesion. Conversely, if it is too long, insufficient bonding can lead to weak interfaces and increased risk of warping or cracking [2], [3].

Accurately predicting how layer time adjustments affect temperature distribution is essential for maintaining print quality in LFAM. During printing, anomalies such as excessive warping or poor adhesion may indicate suboptimal layer time settings. In such cases, adjustments are necessary, but determining the optimal modification requires understanding its impact on the thermal profile of the printed part [4]. Digital Twin technology provides a powerful solution by creating a real-time virtual replica of the printing process, allowing continuous monitoring of temperature evolution under different layer time settings. By analyzing how temperature distribution changes with varying layer times, a DT helps operators detect thermal anomalies, assess process stability, and make informed adjustments to optimize printing parameters and ensure high-quality fabrication.

Numerous studies have explored Digital Twin applications in LFAM, which can be categorized into Physics-Based and Data-Driven approaches. The Physics-Based DT relies on numerical simulations and analytical models to predict thermal behavior. In paper [5], a 3D Finite Element Analysis (FEA) simulation is developed to accurately predict thermal profiles based on printing parameter adjustments. Liu et al. introduced a reduced physics-based model that incorporates a 1D physics-based heat transfer model as a substitute for FEA simulations, enabling offline temperature prediction and layer time optimization [6]. In contrast, Data-Driven DT leverages experimental data and machine learning for faster, more adaptive temperature predictions. Wang et al. used real-time thermal imaging and a linear regression model to efficiently predict surface temperature at a single location, enabling online control [7]. Xie et al. introduced a Transformer-based model trained on historical data to predict temperature distributions across the entire print surface, capturing spatial-temporal dependencies for offline optimization [8]. Moreover, numerous studies in additive manufacturing leverage the advantages of both physics-based modeling and data-driven approaches by developing hybrid physics-informed neural networks (PINNs) to simulate the thermal behavior of the printing process [9].

However, a key limitation of these approaches is that their Digital Twin models generate static one time predictions without the ability to dynamically update thermal profile predictions based on real time printing strategy changes. To overcome this limitation, we leverage Generative AI to develop an adaptive Digital Twin framework that dynamically updates predictions when printing parameters, such as layer time, are adjusted. Diffusion Models, particularly Denoising Diffusion Probabilistic Models (DDPMs) [10], have demonstrated strong generative capabilities in image and video synthesis, gradually transforming random noise into structured outputs through iterative denoising. These models have been widely applied in image generation and extended to video synthesis by incorporating temporal consistency, making them well-suited for predicting thermal evolution in LFAM.

Recent studies have explored using Variational Autoencoders (VAEs) [11] to reduce the computational cost of diffusion models by mapping images into lower-dimensional latent representations. Meanwhile, Transformer-based diffusion models [12], [13], [14], [15] have been shown to outper-

Lu Liu, Haoyang Xie, and Feng Ju are with the School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281 USA.

Dylan Hoskins and Kyle Rowe are with Haddy, St. Petersburg, FL 33705 USA.

Please send all correspondance to Dr. Feng Ju at fengju@asu.edu.

form traditional U-Net architectures [16] in video generation by effectively capturing spatial and temporal dependencies.

Inspired by these advancements, we propose a Video Diffusion Transformer (VDT) based Digital Twin, which leverages past thermal frames and layer time settings to generate predictive thermal videos when printing parameters are modified. Unlike conventional methods that provide static, one-time predictions, our approach introduces adaptability, allowing the model to dynamically update temperature forecasts in response to layer time adjustments. By conditioning on observed thermal states and their associated layer time settings, the VDT-based Digital Twin can generalize across different configurations and synthesize updated future thermal states, providing a visually intuitive and data-driven approach to temperature prediction. By leveraging Generative AI, this adaptive Digital Twin framework enhances LFAM process flexibility, enabling dynamic optimization and informed decision-making. The ability to generate high-fidelity thermal video sequences provides a more intuitive visualization of thermal evolution, making it a valuable tool for real-time process monitoring and control.

The remainder of this paper is structured as follows. Section II presents the problem description. Section III details the proposed methodology, introducing the Video Diffusion Transformer-based Digital Twin framework. Section IV provides the experiment evaluation and results analysis, demonstrating the effectiveness of our approach. Finally, Section V concludes the paper and discusses potential future research directions.

## II. PROBLEM DESCRIPTION

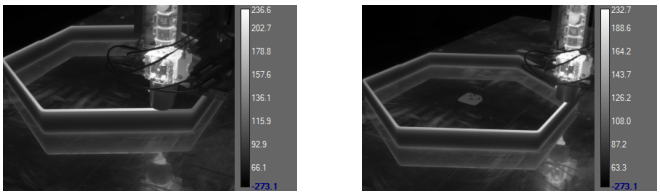
In this study, we utilize the CEAD AM Flexbot for LFAM. The printing material used is 30% glass fiber reinforced PETG, selected for its high strength and thermal stability. The experimental settings are detailed in Table I.

To achieve the Video Diffusion-based Digital Twin framework outlined earlier, we conduct experiments with two different layer time settings: 3 minutes and 4 minutes, while keeping all other parameters identical. The test geometry is a hexagon. The detailed geometric parameters are also provided in Table I

TABLE I: Experimental setting.

| **Parameter** | **Setting 1** | **Setting 2** |
|---|---|---|
| Deposition temperature | 200°C | |
| Ambient temperature | 24°C | |
| Hexagon perimeter | 3564 mm | |
| Hexagon height | 365 mm | |
| Bead width | 20 mm | |
| Thickness | 5 mm | |
| Layer time | 180 s | 240 s |
| Printer head speed | 20 mm/s | 15 mm/s |

Throughout the printing process, we employ a FLIR thermal camera with a resolution of $256 \times 320$ pixels to capture continuous thermal images, recording the entire temperature evolution. Each pixel in the FLIR images corresponds to a specific temperature value, allowing for a direct mapping between grayscale intensity and temperature distribution. This enables precise thermal analysis, which is crucial for constructing and validating our Video Diffusion-based Digital Twin. Figure 1 illustrates sample FLIR images captured during the experiments.



(a) 3-minute layer time. (b) 4-minute layer time.

Fig. 1: Thermal images taken by FLIR™ camera.

To achieve an adaptive Digital Twin, our goal is to predict future thermal distributions based on the current printing settings and previously captured thermal frames. Specifically, we aim to model how changes in printing parameters, such as layer time, influence the future frames of the printing process. Let $F_t$ represent a single thermal frame captured at time t. Given a sequence of past frames $\{F_{t-n}, \ldots, F_{t-1}, F_t\}$, we seek to generate future frames $\{F_{t+1}, F_{t+2}, \ldots, F_{t+m}\}$ under different printing conditions.

To achieve this, during the training phase, we use past frames from a specific layer time setting and incorporate a corresponding label that encodes this setting into the VDT. This guides the model to generate future frames that match the thermal evolution under that configuration. By conditioning the model on layer time settings, we enable the VDT-based Digital Twin to learn the relationship between layer time and temperature evolution, allowing it to generalize to new layer time adjustments. This capability is crucial for evaluating whether the model can adaptively generate realistic temperature distributions when layer time is modified. The detailed implementation of this framework is described in Sec. III.

## III. PROPOSED METHODOLOGY

In this paper, we leverage the strong performance of the Video Diffusion Transformer (VDT) architecture in predicting future video frames, and propose incorporating an additional instructional condition to modulate prediction features, thus enabling enhanced video generation capabilities for achieving digital twins in LFAM.

### *A. Diffusion Model*

The proposed video generation architecture employs a denoising diffusion probabilistic model (DDPM) [10] as its backbone. In the forward process of the DDPM model, Gaussian noise is progressively added over timesteps $T$ to the original clean image $x_0$. The timesteps $T$ are also referred to as the noise levels. the noised image $x_t$ at noise level $t$ can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon \tag{1}$$

where $\bar{\alpha}_t$ is a constant hyperparameter related to the predefined noise variance, and $\varepsilon \sim N(0, I)$ represents Gaussian

noise. The backward process of DDPM is used to iteratively sample a less noisy image $x_{t-1}$ from the posterior distribution $p_\theta(x_{t-1}|x_t)$ by training a function approximator $\varepsilon_\theta$ to predict noise $\varepsilon$ from $x_t$. After simplification, the denoised image $x_{t-1}$ is formulated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t,t)\right) + \sigma_t z \tag{2}$$

where $\varepsilon_\theta(x_t,t)$ is the predicted noise from a neural network model at noise level $t$, $\sigma_t$ is a noise-related term and $z \sim N(0,I)$ is an additional Gaussian noise component. The term $\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t,t)$ refines the estimated noise and removes it from $x_t$. The stochastic term $\sigma_t z$ introduces controlled randomness into the reverse process, aiding in improved sample diversity. The mean squared error (MSE) loss function is used to minimize the difference between the true noise $\varepsilon$ added in the forward process and the noise predicted by the model $\varepsilon_\theta(x_t,t)$. The loss function is formulated as:

$$\mathscr{L}(\theta) = \mathbb{E}_{x_0,\varepsilon,t}\left[\|\varepsilon - \varepsilon_\theta(x_t,t)\|^2\right] \tag{3}$$

In general, the U-Net architecture is the most commonly used model for predicting noise in DDPM for image generation. However, recent research has demonstrated that transformer-based architectures outperform U-Net in generating images within the DDPM framework. Therefore, in this paper, we propose a digital twin model utilizing a transformer-based diffusion architecture to generate future video clips based on adjustment prompts. The details of the proposed architecture are discussed in the following section.

### *B. Video Diffusion Transformer based Digital Twin*

During the printing process of LFAM, monitoring the surface layer temperature is crucial, as it directly impacts the final product quality. If anomalies are detected in the layer temperature, adjustments to the printing parameters become necessary. To evaluate the appropriateness of these adjustments, we aim to generate potential changes in the printing process through video generation. To achieve a preliminary digital twin for LFAM, we define two conditions for constructing the diffusion-based generative model. First, to ensure accurate video prediction and temporal consistency between generated frames, past video clips are used as a conditioning factor. Second, since layer time adjustments directly influence temperature variations, they are incorporated as the second conditioning factor. A longer layer time results in faster temperature cooling, whereas a shorter layer time leads to slower cooling of the printing surface. By integrating these conditions, the preliminary digital twin model can generate future video sequences that reflect the impact of printing adjustments, aiding in process optimization.

Building upon the effective strategies employed in ViT [13], DiT [14], and VDT [17], the proposed architecture in this work retains most of these beneficial components while being modified to simultaneously incorporate two conditioning factors. In the diffusion model, using the original pixel size as input demands substantial computational resources, which becomes particularly costly for our digital twin application, where generating video clips is required. Therefore, each frame $F_t$ in the original input is projected into the latent representation space $Z_t$ using a pre-trained convolutional VAE model, which effectively reduces dimensionality while preserving essential spatial features [11]. The pre-trained convolutional VAE model applied in this paper downsamples frame size by a factor of 8.

During the forward process of the proposed transformer-based DDPM, Gaussian noise is added to the latent representation sequences, which have a tensor shape $(T,C,H_z,W_z)$, where $T$ represents the number of sequential frames, $C$ denotes the latent feature dimension, and $H_z = H/8$ and $W_z = W/8$ correspond to height and width of the encoded frames, respectively.

In the backward process of the diffusion model, the denoising is performed through a sequence of transformer blocks. The architecture of the transformer block, illustrated in Fig. 2, processes three embedded input sequences: the noise timestep embedding, the embedding of encoded sequential frames, and the layer time embedding. To represent the encoded sequential frames as tokens for the transformer architecture, each frame is divided into patches of size $p \times p$ and linearly transformed into a tensor of shape $(T' \times D)$, where $T' = H_z W_z / p^2$ is the number of tokens, and $D$ denotes the hidden dimension size. Then, the sequence of tokens is embedded using spatial and temporal positional embeddings derived from the sine-cosine approach.

In the design of the transformer architecture, the scale parameter $t_{scale}$ and the shift parameter $t_{shift}$ derived from the noise timestep embedding, are incorporated into each block through adaptive layer normalization (adaLN):

$$adaLN(h,t) = t_{scale} LayerNorm(h) + t_{shift} \tag{4}$$

where h denotes the hidden state of a block in transformer. As described in VDT, the token concatenation method used to integrate our first condition—sequential past frames—not only demonstrates superior consistency in predicting future videos but also enhances model convergence speed. Following this method, the sequential past frames $\{Z_{t-n},...,Z_{t-1},Z_t\}$ are concatenated with the noised future sequential frames $\{Z_{t+1},Z_{t+2},...,Z_{t+m}\}$, where $n$ represents the number of past frames and $m$ denotes the number of future frames to be generated. The resulting sequentially integrated tokens are then fed into the transformer as input, first passing through a multi-head temporal attention mechanism, followed by a multi-head spatial attention mechanism, both utilizing the multi-head self-attention approach. The temporal attention block captures dependencies and correlations across frames, enabling the model to understand motion dynamics over time. Meanwhile, the spatial attention block focuses on positional relationships within each frame, preserving structural consistency and spatial details.

With the objective of developing a preliminary digital twin for LFAM, layer time adjustments for the next printing step are considered as the second prompt, enhancing the model's ability to generate more diverse future videos.

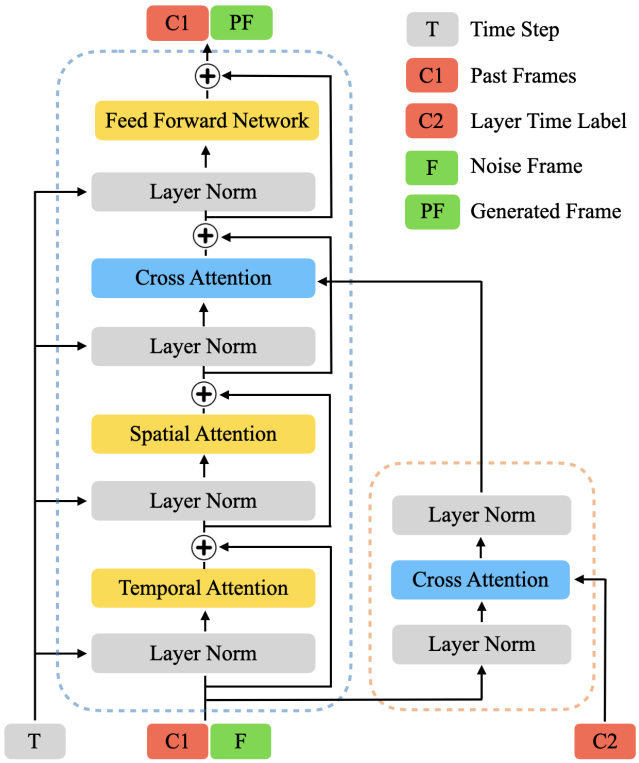



Fig. 2: The framework of the proposed video diffusion transformer based thermal video generation.

Given the limited number of available printing cases, the layer time corresponding to future videos can be intuitively labeled as a class. We propose integrating the embeddings of these prompts into the transformer architecture using two cross-attention blocks. In the first cross-attention block, the attention score is computed between the embedding of encoded sequential frames (serving as the query) and the layer time embedding (acting as the key-value pair) [12]. This setup enables the sequence of frames to selectively focus on relevant aspects of the layer time information, helping it understand how different layer time conditions influence the generated frames. In the second cross-attention block, the hidden state, after being passed through temporal and spatial attention mechanisms (as shown in Fig. 2), is further processed through cross-attention with the output from the first cross-attention block. This step reinforces the integration of layer time information with the learned temporal and spatial features, enabling the model to generate future video frames that accurately reflect the impact of printing adjustments.

Finally, a feedforward network utilizing the MLP is applied to transform the generated representations into a sequence of concatenated frames $\{Z_{t-n},...,Z_t,Z'_{t+1},...,Z'_{t+m}\}$ where $Z'$ represents the generated denoised future frames in the latent space. Once all denoising iterations have been completed through the sequence of transformer blocks, the generated clean future sequential frames are converted back to their original frame size using the VAE decoder.

### C. Training

In this paper, our proposed Video Diffusion Transformer-based digital twin model is trained in two different model sizes, varying in the depth of the transformer blocks, hidden dimension size, and the number of heads in the multi-head attention blocks, as shown in TABLE II. The patch size for dividing frames is set to 2 in both configurations. Due to CUDA memory constraints, the small model is trained in parallel using two NVIDIA A100 GPUs for 28 hours, stopping upon completion of the designated epochs. The large model is trained in parallel using four NVIDIA A100 GPUs for 36 hours, terminating upon reaching the time limit.

TABLE II: Size of the transformer based diffusion model on video generation [14], [17].

| Model | Depth | Hidden dimension | Heads | Patch |
|---|---|---|---|---|
| Small | 12 | 384 | 6 | 2 |
| Large | 28 | 1152 | 16 | 2 |

There are two datasets of hexagonal geometry printing recorded during the LFAM manufacturing process. The first case is printed with a 3-minute layer time, while the second case is printed with a 4-minute layer time. During training, we use 3000 sequential frames from each dataset. Each sequence consists of 16 frames, where the first 8 frames serve as the conditional past frames, and the remaining 8 frames are the generated future frames. The input sequences are constructed using a moving time window with a stride of 1 frame and are labeled with the corresponding layer time, which serves as the second prompt in the proposed model.

For each frame in the video clip, the original image resolution $F_t$ is $256 \times 320$. To leverage the most accurate results from the pre-trained VAE model, each frame is cropped to $256 \times 256$. Since the video clips are recorded using an IR camera, where each pixel value represents the actual temperature, each frame is normalized before being fed into the VAE encoder. In this work, the downsampling rate of the VAE encoder is set to 8, meaning that the input frames are projected into latent representations $Z_t$ before embedding, resulting in a shape of $32 \times 32$.

During the training process, a learning rate warm-up is applied for the first 5 epochs, starting at $1 \times 10^{-5}$, to stabilize optimization and prevent large gradient updates at the beginning of training. Afterward, a cosine annealing learning rate schedule is used to gradually reduce the learning rate, helping the model converge smoothly. No weight decay is applied to preserve the learned representations without additional regularization constraints. In the backward process of the diffusion model, 500 diffusion steps are performed. The mean squared error (MSE) loss function is used to measure reconstruction accuracy, while the AdamW optimizer is employed to efficiently update model parameters and find the optimal values.

An image view of the video generation results is presented in Fig. 3. In the sequential images, the first 8 frames represent the ground truth past frames, while the last 8 frames correspond to the generated future frames. The extruder

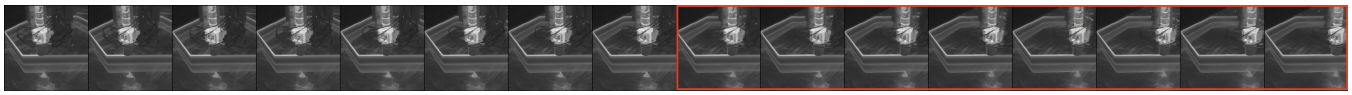


Fig. 3: Image view of the video generation.

moves from left to right, exhibiting consistent motion without noticeable blur. The temperature variations along the printing path are described through changes in luminance, where regions near the extruder display the highest luminance. This observation indicates the model's ability to accurately generate the temperature cooling process.

## IV. EXPERIMENT AND RESULTS ANALYSIS

We evaluate the inference time of our proposed Video Diffusion Transformer-based Digital Twin on two model configurations. The small model requires 5 seconds per sequence (8 future frames), while the large model requires 60 seconds per sequence (8 future frames). The significant difference in inference speed reflects the trade-off between computational efficiency and prediction accuracy.

To assess the effectiveness of our approach, we evaluate the generated thermal videos from two perspectives: realism of the generated videos and accuracy of the temperature distribution.

### *A. Realism of the Generated Videos*

To evaluate the quality of the generated video sequences, we use Fréchet Video Distance (FVD), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). These metrics allow us to assess both the temporal consistency of the video and the accuracy of individual frames.

FVD is a widely used metric for evaluating video generation models. It measures the distributional difference between real and generated videos in a feature space, capturing both spatial quality and temporal coherence. A lower FVD score indicates a smaller discrepancy between generated and real videos, meaning the model produces more realistic motion and frame transitions. In our evaluation, we compute FVD using VideoGPT, which effectively extracts high-level video representations for comparison.

While FVD focuses on overall video quality, PSNR and SSIM are used to evaluate the accuracy of single-frame generation. PSNR measures pixel-wise similarity between generated and real frames, with higher values indicating better fidelity. However, it does not account for perceptual differences that the human eye may notice. To complement this, SSIM is used to assess structural similarity by considering luminance, contrast, and texture patterns. A higher SSIM value suggests that the generated frame maintains a closer resemblance to the ground truth.

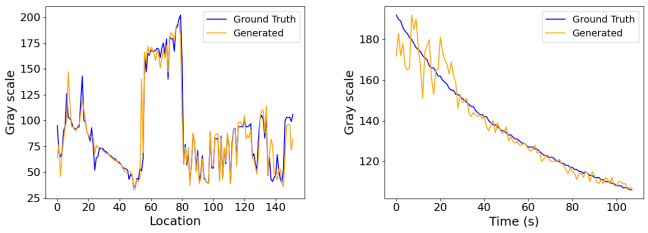
Table III presents the quantitative results for different model configurations. The VDT-Large model consistently achieves lower FVD scores, demonstrating superior video quality with smoother and more realistic frame transitions. It also achieves slightly higher SSIM values, indicating better structural consistency. However, this comes at the cost of significantly higher inference time, making the VDT-Small model a more practical choice for real-time applications.

TABLE III: Performance comparison of VDT-Small and VDT-Large under different layer times.

| Model | Layer Time | FVD ↓ | PSNR ↑ | SSIM ↑ |
|---|---|---|---|---|
| VDT-Small | 3 min | 91.44 | 26.66 ± 8.09 | 0.874 ± 0.105 |
| | 4 min | 258.62 | 27.66 ± 8.09 | 0.922 ± 0.044 |
| VDT-Large | 3 min | 55.26 | 26.20 ± 9.5 | 0.874 ± 0.104 |
| | 4 min | 127.29 | 27.63 ± 7.96 | 0.924 ± 0.043 |

### *B. Accuracy of the Temperature Distribution*

To evaluate the accuracy of the generated temperature distributions, we analyze grayscale intensity in FLIR videos, where grayscale values are directly proportional to temperature. A higher grayscale value corresponds to a higher temperature. While the small model does not perform as well as the large model in terms of generation quality, it has the significant advantage of much faster inference speed, making it more practical for real-time applications. Given this trade-off, we use the small model for this part of the analysis, setting the 3-minute layer time case as input and generating the predicted temperature distribution for an adjusted 4-minute layer time.




(a) Grayscale distribution across spatial positions.

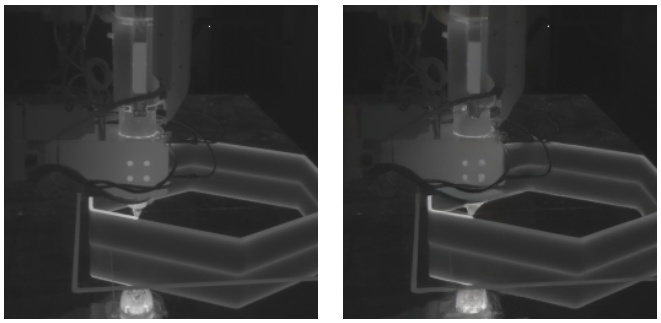
(b) Grayscale profile at a specific location.

Fig. 4: Comparison of grayscale intensity between ground truth and generated results.

Figure 4 provides a comparison between the generated and ground truth grayscale distributions. Subfigure (a) illustrates the grayscale intensity variations across different spatial positions, highlighting how the generated temperature distribution aligns with the real one. The overall trend matches well, though minor deviations exist in certain regions. Subfigure (b) presents the grayscale profile at a specific location over time, showing the cooling process as the grayscale values decrease. The generated results capture the general trend of temperature evolution, demonstrating the small model's ability to maintain temporal consistency.

To further quantify the accuracy of the generated temperature distribution, we compute the mean squared error (MSE)

of grayscale differences, which results in a value of 0.2257. This indicates that while there are some discrepancies, the model effectively captures the underlying thermal patterns, reinforcing the small model's feasibility for fast, approximate temperature predictions in real-time adjustments.



(a) Ground truth image (b) Generated image

Fig. 5: Comparison of images between ground truth and generated results.

Figure 5 presents a visual comparison between the ground truth and generated images. The generated image closely resembles the real captured frame, maintaining the structural integrity and temperature distribution of the printed geometry. The fine details in the grayscale representation are well-preserved, demonstrating the model's capability to generate realistic and coherent predictions.

The high level of visual similarity suggests that the model can accurately predict future states when layer time is adjusted. This capability is particularly useful for real-time process optimization, as it allows for accurate forecasting of temperature evolution when modifying printing parameters. The generated results validate the model's potential in assisting layer time adjustments, ensuring stable thermal conditions and improved print quality.

## V. CONCLUSION

In this work, we propose a Video Diffusion Transformer (VDT) based Digital Twin framework for adaptive thermal prediction in LFAM. Unlike traditional models that provide static thermal estimations, our approach enables dynamic simulation of future temperature distributions when layer time is adjusted. By leveraging past thermal frames, the VDT model generates realistic video sequences that visualize how temperature evolves after modifying layer time.

Our experiments demonstrate that the generated thermal videos align closely with real FLIR recordings, capturing both spatial and temporal thermal dynamics. While the large model provides higher fidelity predictions, the small model achieves significantly faster inference speed, making it more suitable for real-time applications. By using the small model to generate temperature evolution for an adjusted layer time, we showcase the potential of this framework for process optimization and defect prevention in LFAM.

This study highlights the feasibility of using generative AI to enhance the adaptability of Digital Twin systems. Future research could explore further conditioning mechanisms, expand the model to handle additional process parameters, and integrate real-time feedback to improve LFAM process. Our approach advances intelligent, self-adjusting manufacturing workflows capable of dynamically adapting to process variations.

## REFERENCES


[1] C. M. Vicente, M. Sardinha, L. Reis, A. Ribeiro, and M. Leite, "Large-format additive manufacturing of polymer extrusion-based deposition systems: review and applications," *Progress in Additive Manufacturing*, vol. 8, no. 6, pp. 1257–1280, 2023.

[2] S. Fathizadan, F. Ju, F. Wang, K. Rowe, and N. Hofmann, "Dynamic material deposition control for large-scale additive manufacturing," *IISE Transactions*, vol. 54, no. 9, pp. 817–831, 2022.

[3] S. Fathizadan, F. Ju, K. Rowe, A. Fiechter, and N. Hofmann, "A novel real-time thermal analysis and layer time control framework for large-scale additive manufacturing," *Journal of Manufacturing Science and Engineering*, vol. 143, no. 1, p. 011009, 2021.

[4] F. Wang, S. Fathizadan, F. Ju, K. Rowe, and N. Hofmann, "Print surface thermal modeling and layer time control for large-scale additive manufacturing," *IEEE Transactions on automation science and engineering*, vol. 18, no. 1, pp. 244–254, 2020.

[5] E. Jo, L. Liu, F. Ju, D. Hoskins, D. Pokkalla, V. Kunc, U. Vaidya, and P. Kim, "The design of layer time optimization in large scale additive manufacturing with fiber reinforced polymer composites," tech. rep., Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2022.

[6] L. Liu, E. Jo, D. Hoskins, U. Vaidya, S. Ozcan, F. Ju, and S. Kim, "Layer time optimization in large scale additive manufacturing via a reduced physics-based model," *Additive manufacturing*, vol. 72, p. 103597, 2023.

[7] F. Wang, F. Ju, K. Rowe, and N. Hofmann, "Real-time control for large scale additive manufacturing using thermal images," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pp. 36–41, IEEE, 2019.

[8] H. Xie, D. Hoskins, K. Rowe, and F. Ju, "Transformer-based offline printing strategy design for large format additive manufacturing," *Journal of Computing and Information Science in Engineering*, vol. 25, no. 2, 2025.

[9] S. Liao, T. Xue, J. Jeong, S. Webster, K. Ehmann, and J. Cao, "Hybrid thermal modeling of additive manufacturing processes using physics-informed neural networks for temperature prediction and parameter identification," *Computational Mechanics*, vol. 72, no. 3, pp. 499–512, 2023.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

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

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

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[15] C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, and L. Bai, "A survey on long video generation: Challenges, methods, and prospects," *arXiv preprint arXiv:2403.16407*, 2024.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[17] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding, "Vdt: General-purpose video diffusion transformers via mask modeling," *arXiv preprint arXiv:2305.13311*, 2023.