COOL: A Cooling-Aware Point Transformer Framework for Thermal Prediction in Advanced 3D/3.5D IC Packaging

Yao Lu, Zhicheng Guo, Qijun Zhang, Shang Liu, Wenji Fang, Wenkai Li, Zhiyao Xie*
Hong Kong University of Science and Technology
Email: eezhiyao@ust.hk
(*corresponding author)

ABSTRACT

Advanced 3D and 3.5D IC packaging significantly improves integration density but elevates thermal management challenges due to cross-layer heat coupling and complex cooling structures. Traditional solvers deliver high fidelity but are too slow for iterative design flows, while existing learning-based methods either fail to capture inter-die thermal coupling or treat cooling structures as static components, limiting their applicability in real packaging co-design scenarios. In this work, we introduce COOL, a cooling-aware point transformer framework that represents heterogeneous assemblies (dies, interposers, TIMs, heat spreaders) as annotated 3D point clouds embedding geometric, material and power attributes. COOL explicitly encodes geometric boundaries and cooling structures, and introduces a physics-informed boundary condition (PI-BC) loss to enforce thermal consistency at material interfaces and cooling boundaries. Extensive experiments demonstrate that COOL achieves a remarkable 2.4% NMAE on our constructed benchmark of multi-package thermal designs, substantially outperforming existing learning-based approaches while providing over 15.7× speedup compared to commercial FEM solvers.

ACM Reference Format:
Yao Lu, Zhicheng Guo, Qijun Zhang, Shang Liu, Wenji Fang, Wenkai Li, Zhiyao Xie*, Hong Kong University of Science and Technology, Email: eezhiyao@ust.hk, (*corresponding author). 2026. COOL: A Cooling-Aware Point Transformer Framework for Thermal Prediction in Advanced 3D/3.5D IC Packaging. In *63rd ACM/IEEE Design Automation Conference (DAC '26), July 26–29, 2026, Long Beach, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3770743.3804229

1 INTRODUCTION

Advanced 3D and 3.5D IC packaging technologies enable high integration density and performance, but also introduce severe thermal coupling and non-uniform heat dissipation challenges. Accurate thermal analysis is essential for ensuring performance and reliability. However, conventional finite-element method (FEM) or finite-difference method (FDM) solvers such as COMSOL, Cadence Celsius, and ANSYS Icepak are computationally expensive, often requiring hours to simulate a single power trace. In addition, these solvers require substantial manual effort for geometric modeling and meshing. This makes them impractical for iterative design space exploration or layout optimization automatically.

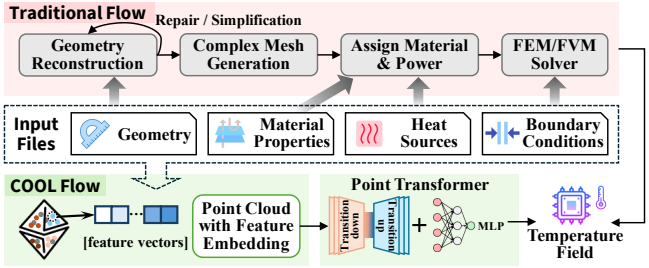


Figure 1: Comparison between the traditional thermal workflow and the proposed COOL framework. Traditional FEM/FVM workflows require manual preprocessing for geometry, meshing, and boundary setup. COOL encodes all inputs into a unified point-cloud embedding and uses a point transformer for direct temperature prediction, enabling fast and automated thermal analysis.

In realistic 3D/3.5D packaging, multiple active dies are vertically stacked and thermally coupled through inter-die interfaces [9, 15] such as thermal interface materials (TIMs), through-silicon vias (TSVs), and micro-bumps. At the same time, chiplets often differ in thickness and material composition, resulting in heterogeneous and highly anisotropic heat transfer paths. These cross-layer interactions and geometric variations jointly determine the temperature field within the package, making accurate modeling far more complex than in single-die systems [4, 6, 12]. Recent learning-based thermal models [2, 3, 7, 8, 14, 17, 19] have achieved promising speedup over traditional FEM solvers. However, they are inadequate for capturing complex thermal dynamics in advanced packaging. These approaches face serious limitations in modeling inter-die thermal coupling and adapting to design-dependent cooling paths, restricting their applicability to real 3D/3.5D IC scenarios as follows.

First, existing models **fail to capture inter-die thermal coupling across multiple stacked layers**. For example, DeepOheat [8] and DeepOheat-v1 [17] learn temperature distributions from 2D or volumetric power maps only within a single active die, their formulations do not model heat propagation between vertically stacked chiplets. Similarly, ASRR-PINN [19] assumes a uniform power map at one cross-section and does not vary power across layers, over-simplifying treating the system as a single-die structure. Such simplifications fail to capture the vertical conduction and lateral spreading effects that dominate real 3D IC thermal behavior.

Second, current learning frameworks **treat cooling structures as static fixed conditions rather than flexible design elements with variation**. Both DeepOHeat models [8, 17] and ASRR-PINN [19] assume fixed boundary conditions, ignoring the influence of cooling structures such as heat spreaders or TIMs. SAU-FNO [7]

DAC '26, July 26–29, 2026, Long Beach, CA, USA

ACM ISBN 979-8-4007-2254-7/2026/07
https://doi.org/10.1145/3770743.3804229

makes progress by explicitly modeling boundary constraints and include cooling-related layers like heat sinks or TIMs. However, its formulation treats these components as pre-defined, hard-coded layers with fixed material properties and thickness. Such rigid modeling cannot capture the design-dependent variations commonly found in advanced packaging, where TIM thickness and chiplet stacking geometry may all be adjusted dynamically during co-design. As a result, existing approaches fail to generalize across packaging configurations with heterogeneous structural variability.

To overcome these challenges, a more flexible geometric representation is needed to model the complex, heterogeneous structures of 3D/3.5D packages, along with a corresponding model capable of rapidly capturing information from this new data format.

In this work, we propose **COOL**, the first cooling-aware data-driven framework that jointly learns heat transfer across chips, packages, and cooling components in Figure 1. We propose two key techniques: **data format** and **point transformer model**. ❶ COOL represents the heterogeneous advanced packaging structures by point cloud, which can naturally encode spatial and material heterogeneity at arbitrary resolutions. Each point can represent a physical location with attributes such as material type, power distribution, or boundary conditions, enabling modeling of thermal behavior across irregular geometries and die interfaces. ❷ COOL applies transformer-based attention mechanisms on this point-cloud formulation as the foundation of our thermal learning framework. To incorporate more physical information, we customize COOL to be aware of the boundary conditions. We use the point-based neural operators [16, 18], which demonstrate strong capability in learning spatially continuous physical fields directly from point sets, achieving both geometric flexibility and data efficiency.

In summary, by representing the 3D thermal domain as a point cloud and applying transformer-based attention mechanisms, COOL captures complex geometric and material interactions efficiently. This unified formulation enables efficient and generalizable temperature prediction across diverse 3D packaging configurations. **Our key contributions are summarized as follows:**

- **COOL** is the first learning-based thermal prediction framework that explicitly models the coupled heat transfer among chips, packages, and cooling structures in realistic 3D and 3.5D IC packaging. This formulation captures the design-dependent effects of TIMs, TSVs, and heat spreaders, which are overlooked by existing grid-based or single-die approaches.
- We propose a **point-cloud-based learning methodology** that represents the entire thermal domain as a collection of spatial points, each annotated with its material property, structural domain type, and power activity.
- We introduce **physics–informed boundary conditioning (PI–BC)** in our learning framework, which injects boundary physics into training, enforcing boundary-conditioned heat transfer and inter-die coupling to enhance accuracy at material interfaces and hotspot regions.
- We construct diverse thermal datasets covering multiple 3D/3.5D packaging configurations. Comprehensive experiments demonstrate that COOL achieves fast convergence and a speedup of 15.7× over commercial FEM solvers, with only an NMAE of around 2.4%, which is significantly lower than the baselines on the testing set.

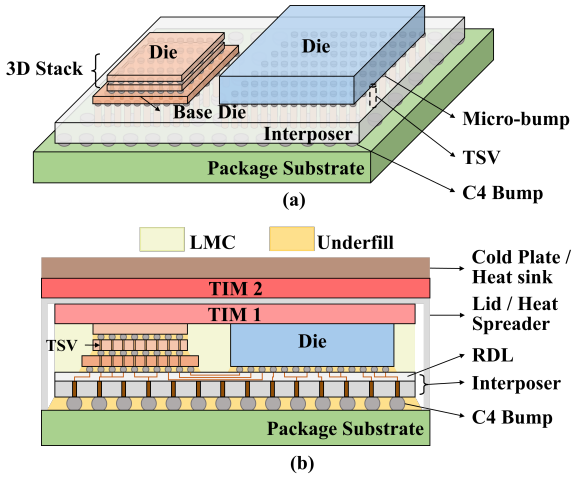


Figure 2: The structure of 3.5D integrated chips, a fusion of 2.5D and 3D-ICs. (a) is a 3D schematic view showing the heterogeneous integration of multiple chiplets on a silicon interposer. (b) Cross-sectional view illustrating CoWoS-based packaging with thermal management structures, including heat spreader, TIM, and heat dissipation devices.

2 BACKGROUND

In this section, we review background on advanced IC packaging and learning-based thermal prediction. We first discuss modeling challenges in 2.5D, 3D, and 3.5D stacks, then summarize prior approaches for incorporating boundary conditions in thermal models.

2.1 Advanced Packaging Modeling

Advanced packaging technologies such as 2.5D and 3D integration have emerged as key enablers for continuing system scaling beyond the limits of traditional monolithic SoCs. In particular, **3.5D integration** combines the advantages of both 2.5D interposer-based and 3D stacked architectures, enabling heterogeneous integration of multiple high-performance chiplets on silicon interposers with vertical interconnects for high bandwidth and low latency.

Because COOL is explicitly designed for these practical 3D/3.5D configurations, it is necessary to model the geometric and material complexity inherent to such packages. Figure 2 illustrates a representative 3.5D structure, where chiplets are bonded onto a silicon interposer through micro-bumps, and heat transfer paths involve multiple materials such as silicon, underfill, die attach, TIMs, and heat spreaders. Accurately capturing these structures requires detailed 3D representations, yet meshing fine-pitch objects such as C4 bumps and TSV arrays is computationally prohibitive.

To build reliable thermal models for learning, we adopt physically meaningful geometric abstractions that retain thermal behavior while remaining simulation-friendly. C4 bumps are approximated as **regularized cubic** primitives with equivalent thermal conductivity, ensuring consistent heat-flow representation without prohibitive meshing cost. Similarly, dense TSV arrays are handled through **area-normalized material homogenization**: instead of meshing thousands of individual TSVs, we compute an effective anisotropic conductivity based on TSV density within each chiplet region. These techniques allow us to simulate real 3.5D structures with high fidelity and produce temperature fields.

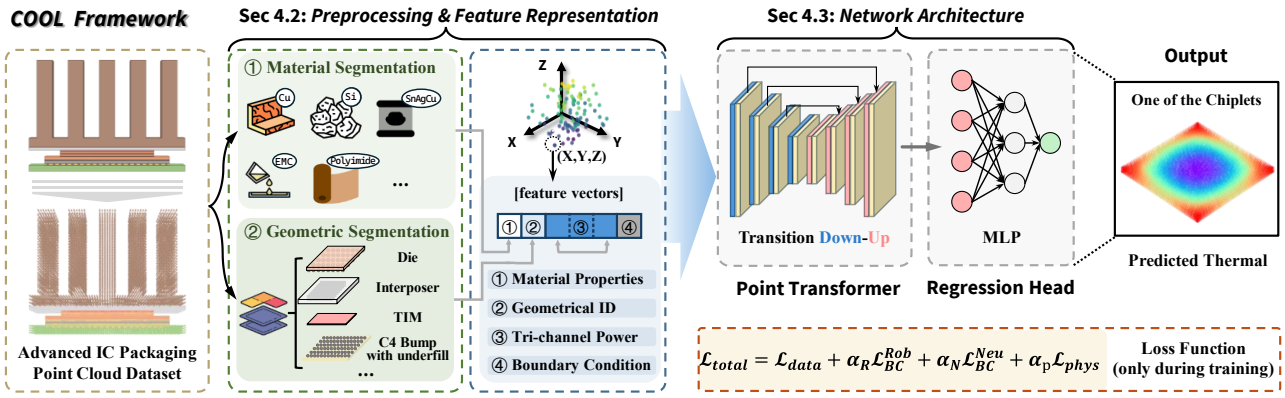


Figure 3: The framework of COOL. Point clouds representing heterogeneous 3D/3.5D IC assemblies are preprocessed with material and structural features and processed by a boundary-aware point transformer. A regression head predicts per-point steady-state temperatures, while the training loss enforces data fidelity, boundary conditions, and physics-inspired consistency.

2.2 Learning-based Thermal Prediction

Learning-based thermal prediction has emerged as a promising approach for accelerating thermal analysis in complex IC stacks. A central challenge is accurately incorporating **boundary conditions** (e.g., Robin, Neumann) and **material transitions**, which strongly influence heat spreading in 3D/3.5D packages.

Prior studies [5, 7, 13, 14, 19] have explored various mechanisms for embedding boundary information into neural models. For example, NeuralMesh [5] enriches the network input by decomposing the 3D structure into material-transition layers and appending per-plane tensors that encode material properties, heat sources, and boundary attributes. ASRR-PINN [19] enforces analytic satisfaction of 3D thermal constraints through a dedicated boundary-conditioning module. Despite these efforts, enforcing boundary physics in highly heterogeneous IC stacks remains challenging due to complex material interfaces and heat-dissipation paths.

These limitations motivate our design of COOL. We adopt a **hierarchical point-based transformer architecture** that directly processes 3D point clouds of advanced packages, avoiding grid quantization error. To ensure physical correctness, COOL integrates a scalable **physics-informed boundary condition (PI-BC)** scheme that incorporates Robin and Neumann interactions directly into the training objective. This combination yields a cooling-aware framework suited for realistic 3D/3.5D IC thermal prediction.

3 PROBLEM FORMULATION

Each 3D/3.5D IC design is represented as a point set and the task is to learn a mapping that outputs the steady-state temperature.

$$P = \{(\mathbf{x}_i, \mathbf{f}_i)\}_{i=1}^{N} \rightarrow (\hat{T}_1, \ldots, \hat{T}_N) \tag{1}$$

where N is the number of the points in a design, $\hat{T}_i$ is the temperature of the point i, $\mathbf{x}_i \in R^3$ are the coordinates and $\mathbf{f}_i$ contains material-domain features, engineered power descriptors, and geometric/boundary indicators as described in Sec. 4.2.

This defines the point-based thermal prediction problem addressed by COOL: predicting physically consistent temperature fields directly from a unified point-cloud embedding.

4 COOL FRAMEWORK

This section introduces the architecture and design of the proposed COOL framework as illustrated in Figure 3. Since we are the first to formulate thermal prediction for advanced IC packaging using point clouds, we develop a dedicated data preprocessing (in Section 4.2) and a corresponding point transformer (in Section 4.3) backbone that unifies chip and cooling structures.

4.1 Overview

The proposed COOL is a **point transformer framework** for steady-state temperature prediction in heterogeneous 3D/3.5D IC packages. Unlike traditional FEM solvers that discretize volumetric geometries onto fixed meshes, COOL directly operates on *point-level representations*, maintaining both spatial fidelity and scalability.

The framework is designed to (1) efficiently scale to designs containing hundreds of thousands of spatial points, (2) explicitly encode heterogeneous material domains and cooling structures, and (3) enhance the learnability of boundary-aware thermal distributions through carefully engineered representations.

4.2 Preprocessing and Feature Representation

The COOL framework begins by transforming each 3D/3.5D IC design into a unified point-based thermal representation. On average, a single design contains around 3×10^5 points, which capture fine-grained structural and material variations across multiple stacked dies, interconnect layers, and packaging materials. This fine-resolution representation enables COOL to model cross-die thermal coupling and heat spreading paths with high fidelity.

Coordinates (X, Y, Z) are normalized to the interval $[0, 1]^3$. Each point includes intrinsic features such as intrinsic thermal-related properties, geometric boundaries, and power dissipation values.

4.2.1 Material-domain Features. Thermal conduction in advanced packaging is governed by interactions among highly heterogeneous materials, such as silicon, copper, underfill, and polymer. However, these materials often exhibit overlapping physical ranges of thermal conductivity (k_{iso}), specific heat (C_p) and density (ρ), which makes it difficult for purely continuous representations to distinguish domains unambiguously.

To address this, COOL adopts a dual-domain encoding scheme. Each material domain is first assigned a categorical identifier that is one-hot encoded and concatenated with its continuous attributes $\{C_p, \rho, k_{iso}\}$. The combined vector is then projected into a material embedding space, enabling the model to learn both domain-specific conduction patterns and cross-material transition behavior. This can help COOL capture heat transfer across different materials.

However, many previous works [5, 19] fail to capture such specific material information in their learning model.

4.2.2 Comprehensive Power Features. Power dissipation (q), serving as the heat source, is extremely sparse and exhibits strong spatial skewness in real 3D IC systems. To capture this distributional complexity, COOL employs an engineered tri-channel power representation: **1) Raw Power:** retains the absolute magnitude of dissipation, preserving direct correspondence to thermal intensity. **2) Binary Flag:** indicates active versus passive regions. It helps guide the model to focus attention on power-generating sites. **3) Log-Transformed Power:** converts long-tailed power distributions. It helps stabilize gradient flow during training.

These comprehensive representations enable COOL to specialize on different source regions, enhancing the model's awareness of local heat density and spatial coherence.

4.2.3 Geometric and Boundary Features. Accurate steady-state thermal prediction strongly depends on how well the model captures geometric complexity and boundary-driven heat transfer behavior. To handle this, COOL augments each point with a set of geometry-aware encodings. First, different geometry domains are one-hot encoded. Then, boundary-type indicators are assigned to points whose coordinates lie on heat-exchange surfaces, including (1) convection boundaries flag (b_i^{conv}) and (2) adiabatic boundaries flag (b_i^{neu}). These features help the network localize boundary conditions that govern heat entering or exiting the system.

Overall, these encoding strategies yield a unified and physics-aware point representation for 3D/3.5D ICs. By embedding material properties, power descriptors, and geometry–boundary cues into each point, COOL provides the essential physical context required by the downstream transformer.

4.3 Network Architecture

COOL adopts a hierarchical point-based transformer architecture. We extend it with thermal-aware sampling and physics-informed constraints to better capture thermal behavior, cross-material conduction, and boundary-driven heat exchange that are critical in 3D/3.5D ICs. The network follows the encoder–decoder paradigm commonly used in geometric deep learning [1].

4.3.1 Encoder with Transition Down. The encoder progressively reduces point resolution with several transition down (TD) layers and point transformer blocks while increasing feature capacity, enabling the network to reason jointly about local conduction patterns and global heat-spreading structures.

① **Transition Down**. Each TD stage consists of sampling, neighborhood grouping, feature aggregation.

Sampling. Standard farthest point sampling (FPS) relies solely on geometry and may overlook thermally important regions such as hotspots and interfaces. COOL incorporates **temperature–aware farthest point sampling (TA–FPS)**, which biases FPS using an importance score that reflects local thermal activity and material transitions. To quantify importance, TA–FPS computes a composite score $s_i \in [0, 1]$:

$$s_i = \alpha_1 \hat{g}_i + \alpha_2 \hat{q}_i + \alpha_3 \hat{b}_i \tag{2}$$

$\hat{g}_i$, $\hat{q}_i$, and $\hat{b}_i$ are normalized temperature gradient magnitude, normalized power density, and boundary indicator, respectively. The classical FPS distance metric D_i is biased as $D_i' = (1 + \gamma s_i)\, D_i$.

This sampling strategy ensures that regions influencing heat flow are retained at coarser resolutions, improving the fidelity of cross-scale thermal reasoning.

Neighborhood Grouping. For each sampled point, a local neighborhood is constructed using k-nearest neighbors (k-NN) [11]. Relative coordinates and local feature differences are concatenated to provide spatial and thermal context for aggregation.

② **Point Transformer Block**. Each TD stage includes a point transformer module that refines local representations using attention mechanisms. Given a point's latent vector and its surrounding neighbors, the module computes attention weights based on learned positional encodings and feature affinity. This allows the network to adaptively prioritize neighbors that influence heat transfer.

4.3.2 Decoder with Transition Up. The decoder mirrors the encoder and gradually restores the prediction to full resolution.

Transition Up (TU). Each TU stage upsamples features from a lower-resolution point set to a higher-resolution one.

Interpolation. Features are propagated using distance-weighted interpolation over a small k-NN neighborhood, ensuring smooth reconstruction across materials and interfaces.

Skip Connections. Encoder features are aligned and concatenated with decoder features to restore fine-grained structure lost during downsampling.

4.3.3 Global Structure and Regression Head. Across the hierarchy, the feature dimension expands from $d = 128$ at the input embedding to progressively higher capacities at deeper layers. At the end of the decoder, a two-layer MLP maps point features to a scalar temperature prediction.

4.3.4 Physics–Informed Boundary Conditioning. Thermal behavior in packaged ICs is strongly influenced by boundary interactions, including heat spreading through heat sinks, convection at exposed surfaces, and continuity of flux across material interfaces.

To capture these effects, COOL introduces PI-BC, which operates jointly with the data supervision loss. The total loss is defined as

$$L = L_{\text{data}} + \alpha_{\text{R}} L_{\text{BC}}^{Rob} + \alpha_{\text{N}} L_{\text{BC}}^{Neu} + \alpha_{\text{p}} L_{\text{phys}} \tag{3}$$

where α_{R}, α_{N}, and α_{p} are coefficients.

Data term: L_{data} is the mean squared error of N pints between predicted $\hat{T}$ and ground-truth T temperatures.

$$L_{\text{data}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{T}_i - T_i \right)^2 . \tag{4}$$

Physics term: Temperature within a homogeneous material domain should remain spatially smooth except near heat sources. To discourage non-physical jumps, we introduce L_{phys}, a Laplacian-based regularization that penalizes deviations between each point and a weighted average of its same-material neighbors.

Boundary condition term: The boundary loss terms L_{BC}^{Rob} and L_{BC}^{Neu} are formulated similarly to the data term, but are selectively applied to points belonging to Robin and Neumann boundaries [10], respectively. For each boundary type, we compute the mean squared error between predicted and ground-truth temperatures.

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 Dataset Construction. We first generate a diverse dataset of 74 heterogeneous 3D and 3.5D IC packaging designs. These designs are constructed by systematically varying key structural parameters listed in Table 1. The parameter set spans multiple aspects of advanced packaging, including: (1) **chip and substrate geometry** (e.g., lateral dimensions $L_{x,y}$, die and interposer sizes) that determine heat–spreading distances; (2) **vertical stacking configurations** such as the number of dies N_{die} and number of stacks N_{stacks}, covering both monolithic dies and multi-die 3D stacks; (3) **interconnect-related thermal paths**, including TSV density f_{TSV} and C4 bump parameters, which influence vertical heat conduction; and (4) **cooling-structure parameters** such as TIM, lid, and heat-sink thicknesses ($L_z^{\text{TIM}}, L_z^{\text{lid}}, L_z^{\text{hs}}$) directly affecting heat extraction efficiency. Sampling each variable within the ranges of Table 1 yields design instances that collectively various 3D/3.5D designs.

For each of the 74 generated designs, we additionally create 10 power-map variants, resulting in a total of **740** simulated thermal scenarios. Steady-state thermal fields are then computed using an FEM solver under realistic boundary and cooling conditions. The resulting temperature distributions are sampled into 3D point clouds, forming the dataset used for training and evaluating **COOL**.

To avoid geometric leakage, splits are performed at the design level. We randomly choose 80% of the designs for training and 20% for testing. Based on the different structures of the designs, there are a total of 58 designs, generating 580 training samples. The remaining 16 designs, with 160 samples, are used for testing.

5.1.2 Preprocessing. To enable batch training and ensure architectural consistency, the original point clouds are resampled to a fixed size of 300,000 points using farthest point sampling (FPS). For each point, we construct the feature vector defined in Section 4.2.

The process includes: (1) normalized 3D coordinates (**3** dimensions); (2) material-domain descriptors, consisting of an **7**-dimensional one-hot material ID and **3** physical properties (k_{iso}, C_p, ρ); (3) **11**-dimensional geometry one-hot encoding and **2**-dimensional boundary indicators, and power-related features.

Concatenating all components yields a 26-dimensional raw feature vector. Before entering the network, this vector is mapped to a 64-dimensional through a feature embedding layer.

5.1.3 Model Setting. COOL follows the hierarchical encoder–decoder architecture in Section 4.3 with fixed-resolution point inputs. The encoder performs three transition down stages, reducing the point set from 300K $\rightarrow$ 150K $\rightarrow$ 75K $\rightarrow$ 37.5K, while expanding feature channels as 64 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 512. Each stage includes TA–FPS sampling and a point transformer block. The decoder mirrors this hierarchy with symmetric Transition Up stages using interpolation-based feature propagation and skip connections.

A lightweight MLP head (64 $\rightarrow$ 64 $\rightarrow$ 1) produces normalized temperature predictions for all points.

5.1.4 Training and Testing. Training uses the AdamW optimizer with initial learning rate 1×10^{-3} and weight decay 1×10^{-4}. A ten-epoch linear warm-up is followed by cosine annealing. Distributed data-parallel training on eight NVIDIA RTX 4090 GPUs supports batch size = 16. The full training schedule runs for 200 epochs and requires approximately 35.6 hours of wall-clock time. During testing, inference is executed on a single RTX 4090 GPU. The predicted normalized temperatures are inverse-transformed to obtain physical temperature fields.

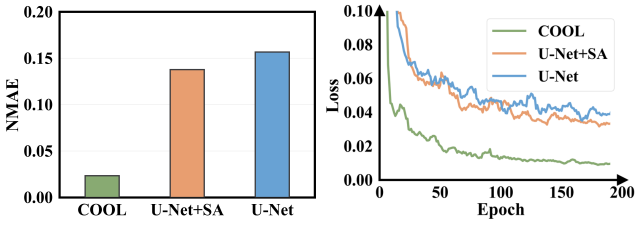
Variable	Description	Range [Unit]
$L_{x,y}$	Substrate lateral size (x-y)	8-16 [mm]
L_z	Substrate thickness	400 [µm]
L_x^{Si}	Interposer lateral size	6-14 [mm]
L_{C4}	C4 bump height	80-120 [µm]
f_{TSV}	TSV density in interposer	5–30% of Si area
N_{stacks}	Number of stacks	1,2,3
N_{die}	Number of stacked dies	2,3,4
$L_{x,y}^{(stack)}$	3D-stacked IC lateral size	4-7 [mm]
$L_z^{(stack)}$	3D-stacked IC thickness	100-250 [µm]
$L_{x,y}^{(mono)}$	Monolithic die lateral size	6-10 [mm]
$L_z^{(mono)}$	Monolithic die thickness	300-800 [µm]
L_z^{TIM}	TIM thickness	100-220 [µm]
L_z^{lid}	Lid thickness	500-800 [µm]
L_z^{hs}	Heat sink thickness	600-1800 [µm]

Table 1: Design variables and parameter ranges used for constructing the 3D/3.5D IC thermal dataset.

5.1.5 Error Metrics. We evaluate prediction accuracy using the **mean absolute error (MAE)** and its range-normalized counterpart, the **normalized MAE (NMAE)**. MAE measures the average absolute difference between predicted and ground-truth temperatures, while NMAE provides a scale-independent error expressed as a percentage of the thermal range. They are defined as

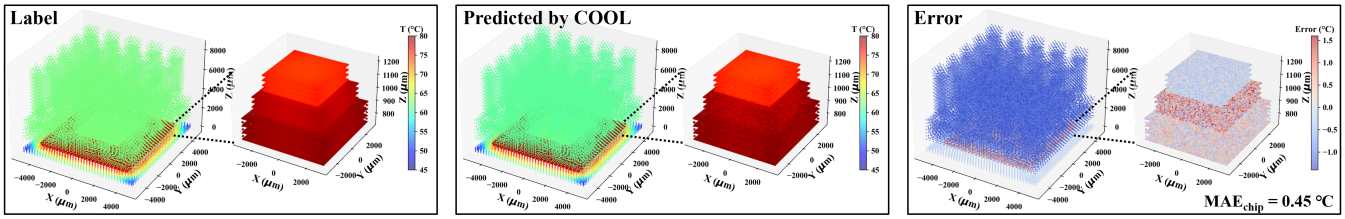
$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|T_i - \hat{T}_i|, \quad \text{NMAE} = \frac{\text{MAE}}{T_{\max} - T_{\min}} \times 100\% \tag{5}$$

where $\hat{T}_i$ and T_i and are the predicted and label temperatures at point i, N is the number of sampled points, and $T_{\max}$ and $T_{\min}$ denote the maximum and minimum temperatures of the design.

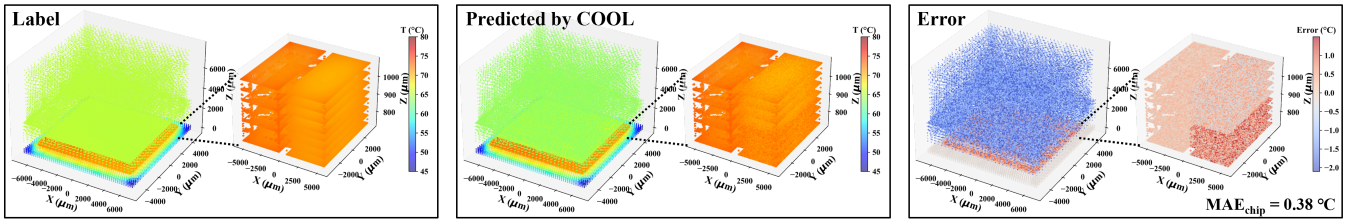


(a) (b)

Figure 4: Comparison among COOL (ours), U-Net, and U-Net+SA. (a) In the evaluation stage, the testing results show that COOL consistently outperforms the two baselines, confirming that our method achieves better overall prediction accuracy. (b) During the training stage, COOL exhibits a faster and more stable convergence in training loss.



(a) The first case study from our designs: $N_{stacks} = 1$, $N_{die} = 3$.



(b) The second case study from our designs: $N_{stacks} = 2$, $N_{die} = 2$.

Figure 5: Three-dimensional point cloud result visualization of temperature for specific designs.

5.2 Baseline

We use commercial FEM tools to generate the temperature field as ground truth. Since SAN-FNO [7] is originally designed for grid-based inputs and cannot directly handle point-cloud data, we construct comparable point-based baselines for comparison. Specifically, we implement a plain U-Net and an enhanced U-Net+SA (U-Net with Self-Attention) following similar design principles. These two models serve as point-based counterparts to SAN-FNO [7], enabling a fair evaluation against our proposed COOL framework.

5.3 Results

In order to validate the effectiveness of COOL, we conducted experiments on heterogeneous IC designs under the experimental settings described in Section 5.1. The aim was to assess both the predictive accuracy and computational efficiency of COOL.

Figure 4 presents a general comparison between COOL and baseline methods during both training and testing. It is evident from the figure that COOL exhibits a consistently lower loss trajectory throughout the process, suggesting more stable learning behavior.

Figure 4a reports testing results on the NMAE metric at *epoch = 200*. COOL consistently outperforms the baselines, achieving an NMAE of approximately 2.4%, while U-Net and U-Net+SA reach 15.6% and 13.7%, respectively. This substantial margin highlights COOL's superior predictive capability and confirms that it can accurately capture the underlying thermal patterns from point-cloud input data. Importantly, this performance gap suggests that traditional point-based and U-Net-based approaches struggle to generalize across the full spatial variability of heterogeneous IC designs, whereas COOL effectively models these complexities.

Figure 4b illustrates the training dynamics of COOL compared to point-based baselines, U-Net and U-Net+SA, over the first 200 epochs. COOL's loss decreases rapidly and approaches convergence around epoch 160, whereas both U-Net and U-Net+SA continue to decline noticeably, indicating slower convergence. This rapid convergence of COOL not only reduces the required training time but also suggests improved optimization stability, potentially leading to more robust performance across different IC designs.

Figure 5 provides a visualization of COOL's prediction performance on two heterogeneous IC designs. The figure includes both full-chip temperature prediction results and detailed temperature distributions over the primary heat sources (chip regions). Figure 5a corresponds to a fully three-dimensional stacked IC, while Figure 5b shows a 3.5D configuration with a separate chip module. COOL accurately captures spatial thermal gradients and temperature patterns in both scenarios, achieving particularly high fidelity in active chip regions with MAEs of approximately 0.45°C and 0.38°C for the 3D and 3.5D designs, respectively.

These visual results further confirm that COOL not only achieves low global error but also delivers precise temperature estimation in critical regions, reinforcing its effectiveness for heterogeneous ICs.

We also provide a rough estimate of runtime. Using a commercial FEM solver as a reference, predicting each design typically requires 189 seconds, excluding manual interventions such as mesh refinement or parameter tuning. COOL completes prediction in 12 seconds per design without any manual intervention, resulting in a speedup greater than 15.7×. The combination of rapid convergence, high predictive accuracy, and low computational cost demonstrates that COOL provides a practical solution for real-time thermal analysis and design verification in heterogeneous ICs.

6 CONCLUSION

In this work, we propose COOL, a cooling-aware point transformer for thermal prediction in advanced 3D/3.5D IC packaging. By representing the thermal domain as an annotated point cloud, COOL explicitly models cooling structures as dynamic design elements. The proposed PI-BC scheme enforces physical consistency at material interfaces and cooling boundaries. Extensive experiments show COOL achieves 2.4% NMAE, substantially outperforming point-based baselines while providing 15.7× speedup over FEM solvers. These results confirm COOL's practicality for thermal-aware co-design in next-generation IC packaging.

ACKNOWLEDGEMENT

This work is supported by Hong Kong Research Grants Council (RGC) CRF-YCRG C6003-24Y, GRF 16216825, and T46-415/25-R. It was partially conducted by ACCESS – AI Chip Center for Emerging Smart Systems, supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

REFERENCES

[1] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine* 34, 4 (July 2017), 18–42. https://doi.org/10.1109/msp.2017.2693418

[2] Liang Chen, Wentian Jin, and Sheldon X-D Tan. 2022. Fast thermal analysis for chiplet design based on graph convolution networks. In *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 485–492.

[3] Liang Chen, Jincong Lu, Wentian Jin, and Sheldon X-D Tan. 2023. Fast full-chip parametric thermal analysis based on enhanced physics enforced neural networks. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 1–8.

[4] Liang Chen, Sheldon X-D Tan, Zeyu Sun, Shaoyi Peng, Min Tang, and Junfa Mao. 2020. A fast semi-analytic approach for combined electromigration and thermomigration analysis for general multisegment interconnects. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 40, 2 (2020), 350–363.

[5] Pengju Chen, Dan Niu, Dekang Zhang, Wenhao Wang, Depeng Xie, Zhou Jin, Wei W Xing, and Lei He. 2025. NeuralMesh: Neural Network For FEM Mesh Generation in 2.5 D/3D Chiplet Thermal Simulation. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–7.

[6] Yinxiao Feng and Kaisheng Ma. 2022. Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*. 121–126.

[7] Zhen Huang, Hong Wang, Wenkai Yang, Muxi Tang, Depeng Xie, Ting-Jung Lin, Yu Zhang, Wei W Xing, and Lei He. 2025. Self-Attention to Operator Learning-based 3D-IC Thermal Simulation. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–7.

[8] Ziyue Liu, Yixing Li, Jing Hu, Xinling Yu, Shinyu Shiau, Xin Ai, Zhiyu Zeng, and Zheng Zhang. 2023. DeepOHeat: operator learning-based ultra-fast thermal simulation in 3D-IC design. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[9] Jiwoo Pak, Sung Kyu Lim, and David Z Pan. 2014. Electromigration study for multiscale power/ground vias in TSV-based 3-D ICs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 33, 12 (2014), 1873–1885.

[10] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. 2021. Learning Mesh-Based Simulation with Graph Networks. arXiv:2010.03409 [cs.LG] https://arxiv.org/abs/2010.03409

[11] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. arXiv:1706.02413 [cs.CV] https://arxiv.org/abs/1706.02413

[12] Yakun Sophia Shao, Jason Clemons, Rangharajan Venkatesan, Brian Zimmer, Matthew Fojtik, Nan Jiang, Ben Keller, Alicia Klinefelter, Nathaniel Pinckney, Priyanka Raina, et al. 2019. Simba: Scaling deep-learning inference with multi-chip-module-based architecture. In *Proceedings of the 52nd annual IEEE/ACM international symposium on microarchitecture*. 14–27.

[13] Zeyu Sun, Weijie Tong, Xiaoning Ma, He Cao, Jianyun Liu, Zhiqiang Li, and Qinzhi Xu. 2025. ChipletEM: Physics-Based 2.5D and 3D Chiplet Heterogeneous Integration Electromigration Signoff Tool Using Coupled Stress and Thermal Simulation. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–7.

[14] Zhenyu Wang, Pragnya Sudershan Nalla, Jingbo Sun, A Alper Goksoy, Sumit K Mandal, Jae-sun Seo, Vidya A Chhabria, Jeff Zhang, Chaitali Chakrabarti, Umit Y Ogras, et al. 2025. HISIM: Analytical Performance Modeling and Design Space Exploration of 2.5 D/3D Integration for AI Computing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2025).

[15] Zhenyu Wang, Jingbo Sun, Alper Goksoy, Sumit K Mandal, Yaotian Liu, Jae-Sun Seo, Chaitali Chakrabarti, Umit Y Ogras, Vidya Chhabria, Jeff Zhang, et al. 2024. Exploiting 2.5 d/3d heterogeneous integration for ai computing. In *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 758–764.

[16] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. 2024. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4840–4851.

[17] Xinling Yu, Ziyue Liu, Hai Li, Yixing Li, Xin Ai, Zhiyu Zeng, Ian Young, and Zheng Zhang. 2025. DeepOHeat-v1: Efficient Operator Learning for Fast and Trustworthy Thermal Simulation and Optimization in 3D-IC Design. *arXiv preprint arXiv:2504.03955* (2025).

[18] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16259–16268.

[19] Zijian Zhou, Min Tang, and Liang Chen. 2025. ASRR-PINN: Adaptive Sub-Regional Random Resampling-Based PINN for Thermal Analysis of 3D-ICs. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–7.