ATMPlace: Analytical Thermo-Mechanical-Aware Placement Framework for 2.5D-IC

Qipan Wang *Student Member, IEEE*, Tianxiang Zhu *Student Member, IEEE*,
Tianyu Jia *Member, IEEE*, Yibo Lin *Member, IEEE*, Runsheng Wang *Member, IEEE*, Ru Huang *Fellow, IEEE*

***Abstract*—Rising demand in AI and automotive applications is accelerating 2.5D-IC adoption, with multiple chiplets tightly placed to enable high-speed interconnects and heterogeneous integration. As chiplet counts grow, traditional placement tools—limited by poor scalability and reliance on slow simulations—must evolve beyond wirelength minimization to address thermal and mechanical reliability, critical challenges in heterogeneous integration.**

In this paper, we present ATMPlace, the first analytical placer for 2.5D-ICs that jointly optimizes wirelength, peak temperature, and operational warpage using physics-based compact models. It generates Pareto-optimal placements for systems with dozens of chiplets. Experimental results demonstrate superior performance: 146% and 52% geo-mean wirelength improvement over TAP-2.5D and TACPlace, respectively, with 3–13% lower temperature and 5–27% less warpage — all achieved ∼10× faster. The proposed framework is general and can be extended to enable fast, scalable, and reliable design exploration for next-generation 2.5D systems.

***Index Terms*—2.5D-IC; Thermo-Mechanical Optimization; Chiplet Placement; Thermal Warpage**

I. Introduction

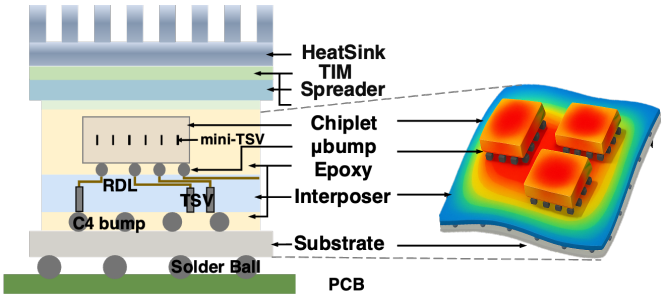


Fig. 1: Layered cross-section (left) and isometric illustration of thermal deformation (right) in 2.5D-IC structure.

This work was supported in part by the National Science Foundations of China (Grant No. 62125401, 62034007), the Natural Science Foundation of Beijing, China (Grant No. Z230002), the Grant QYJS-2023-2303-B, the Beijing Outstanding Young Scientist Program (JWZQ20240101004), and the 111 project (B18001). The preliminary version has been presented at the 43rd IEEE/ACM International Conference on Computer-Aided Design in 2024 [1].

Q. Wang is with the School of Advanced Interdisciplinary and the School of Integrated Circuits, Peking University, Beijing, China.

T. Zhu and T. Jia are with the School of Integrated Circuits, Peking University, Beijing, China.

Y. Lin, R. Wang, and R. Huang are with the School of Integrated Circuits, Peking University, Beijing Advanced Innovation Center for Integrated Circuits, Beijing, China. and Institute of Electronic Design Automation, Peking University, Wuxi, China

Corresponding authors: Yibo Lin (yibolin@pku.edu.cn), and Runsheng Wang (r.wang@pku.edu.cn

TABLE I: Comparison of features of different chiplet placement studies. **Scalability** represents the maximum number of chiplets the method can handle.

Algorithm	Work	Wirelength	Scalability	Thermal aware	Warpage aware
Heuristic	Ho and Chang [2]	Medium	Low	✗	✗
	TAP-2.5D [3]		Low	✓	✗
	TACPlace [4]		Medium	✓	✗
	Zhang et al. [5]		Medium	✓	✗
Enumeration	Osmolovskyi et al. [6]	**Low**	Medium	✗	✗
	SP-CP [7]		Medium	✓	✗
RL	Duan et al. [8]	Medium	Medium	✓	✗
	Deng et al. [9]		Medium	✓	✗
Analytical	Chen et al. [10]	**Low**	Medium	✗	✓
	ATPlace2.5D [1]		**High**	✓	✗
	ATMPlace (ours)		**High**	✓	✓

The continued slowdown of technology scaling has rendered monolithic SoCs increasingly cost-prohibitive, especially in domains that demands both high performance and rapid scalability, such as automotive electronics and AI accelerators. To circumvent these limitations, 2.5D integration has emerged as a practical path forward, assembling pre-verified chiplets on an interposer that delivers high-bandwidth communication, technology heterogeneity, and cost reduction [11]. By enabling reuse and modularity, 2.5D-IC design promises to reshape the semiconductor design paradigm.

However, the effectiveness of 2.5D-ICs fundamentally depends on the placement (floorplan) problem [12].[1] In this stage, the positions and orientations of heterogeneous chiplets must be determined under multiple performance, thermal, and manufacturability constraints.

A wide spectrum of methods has been explored, including heuristic methods (simulated annealing (SA) [2, 3, 13, 14], genetic algorithm (GA) [4, 5]), enumeration-based search [7, 15], reinforcement learning (RL) [8, 9], and analytical optimization [1, 16]. Heuristic methods rely on multi-objective cost functions to co-optimize metrics. However, they require long runtime and often converge to suboptimal layouts. Enumeration-based strategies can deliver good solutions for designs of small instances. Yet they become infeasible when the number of chiplets exceeds ten due to exponential complexity. RL-based approaches offer a data-driven alternative, but they need extensive training data, causing unstable performance. All three categories exhibit limited scalability as chiplet count and diversity increase. For example, heuristic and enumeration methods may require hours for designs with ≥ 10 chiplets, which is impractical for iterative design-space

[1]We use the terminology "placement" and "floorplan" interchangeably to denote the physical arrangement of chiplets.

exploration [10]. In contrast, analytical approaches can handle configurations with up to 60 chiplets in several minutes, as demonstrated in [1].

Moreover, prior work primarily tries to minimize area and wirelength, frequently yielding compact placements with high power density and thermal failure risks. Recent thermal-aware placement studies either embed temperature constraints [13], augment cost functions with worst-case temperature [3], or perform post-optimization refinement [7]. These approaches rely on iterative numerical thermal simulations [17, 18], significantly inflating runtime. Treating temperature as a hard constraint (e.g., limiting peak temperature or chiplet proximity [3, 7]) also narrows the exploration of the broader thermo-design space required by large systems.

Finally, warpage poses a critical reliability challenge in large 2.5D packages with a thin interposer. Industrial reports have highlighted warpage issues in advanced GPU productions [19]. While prior studies analyze and optimize fabrication-stage warpage [10, 20], operational warpage driven by thermal-expansion mismatch and thermal gradients during operation is largely overlooked. The analytical model in [21] targets 2D-ICs but does not capture the geometric and material complexity of 2.5D assemblies.

To conclude, despite a decade of research, current 2.5D-IC placement algorithms remain far from meeting industrial needs. We summarize the features and abilities of previous works in Table I. In this work, we propose an analytical placement framework for 2.5D-ICs with joint thermo-mechanical optimization. To overcome scalability limits, we adopt an orientation-aware analytical placer and couple it with fast physics-based compact models of temperature and warpage. Our main contributions are:

- `ATMPlace`: an analytical chiplet placement framework that jointly optimizes total wirelength and *thermo-mechanical reliability* (temperature and warpage) within a unified flow.
- Two mixed-integer linear programming (MILP) formulations for initialization and legalization that assign chiplet orientations and reduce wirelength while preserving feasibility.
- A physics-based analytical compact model for fast thermal and mechanical evaluation integrated into the optimization loop. It achieves a mean correlation > 0.97 and a $> 8000\times$ speedup for both thermal and mechanical evaluation versus the numerical simulator.

We evaluate our method on large-scale 2.5D-ICs. Under thermo-mechanical-aware placement, it outperforms `TAP-2.5D` and `TACPlace` simultaneously in wirelength (geometric mean improvement: 146% and 52%), peak temperature (reduction: 3% and 13%), and warpage (reduction: 5% and 27%), with $\sim 10\times$ speedup. In wirelength-driven mode, it matches `SP-CP`'s total wirelength within 1% deviation. The rest of the paper is organized as follows: Section II introduces background; Section III details the algorithm; Section IV presents results; Section V concludes the paper.

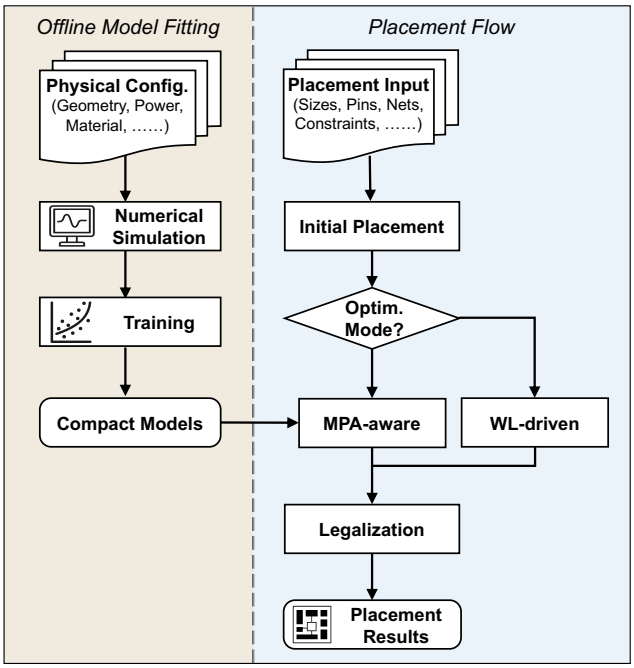


Fig. 2: Illustration of the `ATMPlace` framework, including the compact model fitting and placement flow.

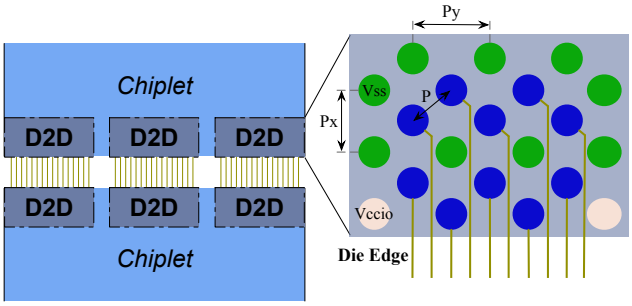


Fig. 3: Illustration of the D2D link between chiplets (left), and a portion of the detailed package bump map of a $\times 16$ module (right). The seashell and green circles are the I/O supply (Vccio) and ground reference (Vss) bump, while the blue circles represent the Tx/Rx bumps, respectively.

II. Preliminaries

In this section, we begin by describing the 2.5D-IC architecture considered in this work in Section II-A. Then we present the thermal and warpage models in Section II-B and Section II-C. Finally, Section II-D formally defines the placement optimization objectives and constraints.

A. 2.5D-IC Configuration

The simplified 2.5D-IC structure studied in this work is illustrated in Fig. 1. It consists of a passive interposer mounted on an organic substrate via C4 bumps. We adopt a transistor-free passive interposer due to its lower cost and design flexibility, while the proposed framework can be extended to active variants. Through-silicon vias (TSVs) pass through the interposer to connect chiplets to external circuits. Chiplets

sit atop the interposer and are modeled as silicon blocks with given power dissipation. They interface with the redistribution layers (RDL) through microbump arrays. Due to their small size, these microbumps are treated as a homogeneous layer with an effective thermal conductivity. The regions between chiplets and microbumps are filled with epoxy, introducing material heterogeneity across the package.

We assume a library of pre-fabricated, modular chiplets is available. Each chiplet includes one or more interface units for data exchange. These interfaces connect to the interposer via microbumps, enabling communication between chiplets and forming complete systems for various applications. To support interoperability among diverse chiplets, standardized interfaces are essential for building a robust chiplet ecosystem. Two primary types of die-to-die (D2D) interfaces exist: serial and parallel. Serial links use a few differential pairs for data transmission, while parallel interfaces employ tens to hundreds of connections. These differ in bandwidth, latency, and packaging requirements, making universal integration across heterogeneous technologies a complex challenge.

For demonstration purposes, we adopt the Universal Chiplet Interconnect Express (UCIe) standard [22] for D2D communication. This choice enables meaningful evaluation and supports future research, though our placement framework is compatible with other interface standards. As shown in the left side of Fig. 3, D2D links connect chiplets through microbump pairs. The right side shows a UCIe reference bump matrix: each signal lane consists of a transmitter (Tx) and receiver (Rx) pair, connected across chiplets. For complex topologies, a hub chiplet may be used to route traffic among multiple chiplets [23].

B. Thermal Model

We analyze steady-state temperature under the worst-case power profile—a standard choice for thermal-aware design. The governing heat conduction equation is:

$$\nabla \cdot (\kappa(\mathbf{r}) \nabla T(\mathbf{r})) = -\mathbf{P}(\mathbf{r}), \tag{1}$$

subject to appropriate boundary conditions. Here, $\kappa(\mathbf{r})$ denotes the spatially varying (and potentially anisotropic) thermal conductivity, and $\mathbf{P}(\mathbf{r})$ is the volumetric power density distribution of the chiplets. Our thermal model naturally supports multiple workload-dependent power profiles: handling M scenarios only requires M independent evaluations to track the worst-case response, without modifying the solver.

Following the common practice [3, 7], we model heat dissipation through a primary path: chiplet $\rightarrow$ TIM $\rightarrow$ heat spreader $\rightarrow$ heatsink $\rightarrow$ ambient air. A secondary path via substrate and PCB is also modeled, and lateral surfaces are assumed adiabatic. We use an air-cooled heatsink with convective resistance of $0.1\mathrm{K/W}$ [18]. Layer sizes, material properties, bump/TSV dimensions follow previous works [3, 24].

Various thermal simulators have been proposed, and mainstream methods encompass numerical and analytical approaches [25]. Numerical methods that mesh the system and solve linear equations form the foundation of a majority of simulators. However, they also demand substantial computation time for high precision, and are hard to derive the gradient information, apart from the adjoint method which brings about high runtime cost. Hence, analytical methods are favored during the design stage, where iterative simulations are required. Such methods can provide rapid solutions through closed-form approximate expressions, either obtaining exact solutions based on simplified models [26] or constructing approximate expressions based on accurate simulation results [27–29]. The latter approach often relies on Green's function, which is based on the impulse response of the system concerning input power given specified initial or boundary conditions.

C. Mechanical Model

Thermal gradients induce mechanical warpage due to the coefficient of thermal expansion (CTE) mismatch across heterogeneous materials (e.g., Si dies, organic substrates, Cu pillars). To assess thermally induced mechanical deformation in the heterogeneous integration stack, we model the out-of-plane displacement $w(\mathbf{r})$ governed by the biharmonic plate equation under non-uniform thermal loading [30]:

$$\nabla^4 w(\mathbf{r}) = \frac{1-\nu}{Eh^3} \nabla^2 \big(\alpha h\, \Delta T(\mathbf{r})\big), \tag{2}$$

where E, ν, and α denote the effective Young's modulus, Poisson's ratio, and coefficient of thermal expansion of the equivalent laminate stack, respectively; h is the effective thickness, and $\Delta T(\mathbf{r}) = T(\mathbf{r}) - T_{\text{ref}}$ represents the local temperature rise obtained from the thermal model in Section II-B. Boundary conditions are assumed to be rigid motion elimination at the package edges as previous [30, 31], and interfacial stresses are captured through homogenized material properties.

The warpage metric is critical for assembly yield and reliability, as it quantifies the global mechanical deformation of the package. It is defined as the peak-to-valley deviation of the out-of-plane displacement field: $\max_{\mathbf{r}} w(\mathbf{r}) - \min_{\mathbf{r}} w(\mathbf{r})$. This formulation enables efficient co-analysis of runtime thermal and mechanical effects.

To simplify finite element analysis, the analytical model proposed in [31] has been widely adopted for its simplicity and computational efficiency on simple bi-layer or tri-layer structures. However, its accuracy is limited when applied to modern heterogeneous packaging with complex material stacks and non-uniform thermal profiles. More recently, Tsai and Wang [32] proposed a warpage model incorporating polynomial expressions with assumed curvature, which is more suitable for warpage estimation in advanced packaging.

D. Problem Formulation

We list key notations in Table II. The goal is to determine the position and orientation (counterclockwise direction) of the chiplets, minimizing the total wirelength, with runtime temperature and warpage satisfying certain conditions meanwhile. Note that the orientation must adhere to certain legal values, denoted by $\Theta = \{\theta^0, \theta^1, \theta^2, \theta^3\}$, where $\theta^{0\sim3}$ are $0^\circ, 90^\circ, 180^\circ, 270^\circ$ respectively.

TABLE II: Some notations used in this paper.

Notation	Meaning
$\mathbb{C}, \mathbb{A}, \mathbb{E}$	Sets of chiplets, connected chiplet pairs, and nets.
$i, j; C_i, C_j$	Index and symbol of a chiplet $\in \mathbb{C}$.
x_i, y_i, θ_i	X and Y coordinates of the center, and rotation angle of the chiplet C_i.
w_i, h_i, t_i	Width, height, and thickness of chiplet C_i.
w_i', h_i'	Width and height of chiplet C_i after rotation.
$\mathbb{B}_i, \mathbb{B}_e$	Set of microbumps belongs to the chiplet C_i or net e.
x_{p_i}, y_{p_i}	X- and Y-offsets from the center of the chiplet C_i for a microbump $p_i \in \mathbb{B}_i$.
w_{gap}	Minimum spacing between two chiplets, set to be $100\mu m$ as in [3].
W, H	Width and height of the placement region (interposer).

In this work, inter-chiplet communication cost is quantified by total wirelength, consistent with prior work [3, 15]. We focus on signal nets (*i.e.*, Tx–Rx pairs in D2D links), assuming power, ground, and I/O connections are handled later. Since these nets are typically two-pin and routed as transmission lines in the interposer, we avoid traditional half-perimeter approximations [33]. Instead, we define the total wirelength $WL(\mathbf{x}, \mathbf{y}, \theta)$ following its exact definition:

$$WL = \Sigma_{e\in\mathbb{E}}\Sigma_{p_i,p_j\in\mathbb{B}_e}\left(\|X_{p_i} - X_{p_j}\| + \|Y_{p_i} - Y_{p_j}\|\right), \quad (3)$$

where p_i and p_j are the two microbumps belonging to net e and

$$\begin{aligned} X_{p_i(j)} &= x_{i(j)} + x_{p_{i(j)}} \cdot \cos\theta_{i(j)} - y_{p_{i(j)}} \cdot \sin\theta_{i(j)}, \\ Y_{p_i(j)} &= y_{i(j)} + x_{p_{i(j)}} \cdot \sin\theta_{i(j)} - y_{p_{i(j)}} \cdot \cos\theta_{i(j)}, \end{aligned} \quad (4)$$

for $i(j)$ representing the corresponding chiplet of microbump $p_i(p_j)$.

To incorporate thermo-mechanical awareness into the optimization framework, we jointly consider thermal distribution and warpage constraints. Let $T_g(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ denote the steady-state temperature field calculated from (1) parameterized by design variables $\boldsymbol{\theta}$, and warpage be the out-of-plane displacement computed from the thermal profile as defined in (2). Specifically, we require that: (i) the temperature at *every* grid point inside the discretized grids $\mathcal{G}$ over the interposer surface satisfies $T(\mathbf{r}; \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) \le T_{\text{th}}$, and (ii) the total package warpage satisfies $\text{Warpage}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) \le W_{\text{th}}$, where T_{th} and W_{th} are reliability thresholds, beyond which performance degradation, electromigration, or thermal runaway may occur. For multiple workloads, the constraints can be conservatively extended by enforcing the limit among all scenarios—a lightweight generalization that scales linearly with the number of profiles.

To avoid overlapping between chiplets, we employ the bell-shaped density function as in [33, 34]. After the placement region (interposer) has been divided into uniform bin grids, the density function $D_b(\mathbf{x}, \mathbf{y}, \theta)$ in bin b is defined as:

$$D_b(\mathbf{x}, \mathbf{y}, \theta) = \sum_{C_i\in\mathbb{C}} \left(D_i \times P_x(b, C_i) P_y(b, C_i)\right), \quad (5)$$

where D_i is the normalization factor, $P_x(b, C_i)$, $P_y(b, C_i)$ are the overlap functions of bin b and chiplet C_i along the x and y directions. Since their plain expressions are neither smooth nor differentiable, the bell-shaped potential function is used to smooth P_x and P_y.

With the preparations above, we can model the chiplet placement as a constrained optimization problem:

$$\begin{aligned} \min \quad & WL(\mathbf{x}, \mathbf{y}, \theta) \\ \text{s.t.} \quad & D_b(\cdot) \le M_b, \quad \text{for each bin } b, \\ & T(\mathbf{r}; \cdot) \le T_{\text{th}}, \quad \text{for } r \in \mathcal{G}, \\ & \text{Warpage}(\cdot) \le W_{\text{th}} \end{aligned} \quad (6)$$

where M_b is the maximum allowable area inside b, defined as: $M_b = t_{max}(w_b h_b)$, where t_{max} is a parameter representing the target density value for each bin, and w_b (h_b) is the width (height) of bin b. Utilizing the quadratic penalty method, it is further translated into a sequence of unconstrained minimization problems:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{y},\theta}\{\mathcal{J}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})\}, \\ &\mathcal{J}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \underbrace{WL(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})}_{\text{wirelength}} \\ &\quad + \underbrace{\lambda_{\text{dens}} \sum_b \left(D_b(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) - M_b\right)^2}_{\text{density penalty}} \\ &\quad + \underbrace{\lambda_T \sum_{\mathbf{r}\in\mathcal{G}} \left[\max\left(0,\, T(\mathbf{r}; \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) - T_{\text{th}}\right)\right]^\gamma}_{\text{thermal penalty}} \\ &\quad + \underbrace{\lambda_W \left[\max\left(0,\, \text{Warpage}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) - W_{\text{th}}\right)\right]^\gamma}_{\text{mechanical penalty}}, \end{aligned} \quad (7)$$

where $\lambda_T, \lambda_W, \lambda_{\text{dens}} > 0$ are tunable weights balancing wirelength, thermal safety, warpage, and density, respectively; $\gamma \in \mathbb{Z}^+$ controls penalty aggressiveness (e.g., $\gamma = 2$ for quadratic, $\gamma > 2$ for stronger discouragement of violations). When all constraints are satisfied ($T \le T_{\text{th}}$, $\text{Warpage} \le W_{\text{th}}$, $D_b \le M_b$), the penalty terms vanish, and the optimizer minimizes wirelength and congestion—yielding a smooth, gradient-friendly objective that prioritizes feasible, high-performance placements. Otherwise, the penalty actively steers the solution toward feasible thermo-mechanical regimes. A value of $\gamma > 1$ yields a superlinear penalty growth with increasing temperature or warpage violations, thereby strongly discouraging high-temperature or high-warpage solutions.

Finally, we formulate the problems in this work as follows.

Problem I (Compact Model Fitting): *Given the physics configuration, including system geometry, material parameters, and power, the goal is to generate a compact thermo-mechanical model that can predict the temperature and displacement distribution as accurately as possible compared to golden simulator.*

Problem II (Placement Optimization): *Given the design specifications and chiplet information—including size, microbump locations, and connectivity—the goal is to supports both wirelength-driven and thermo-mechanical-aware objectives, and can efficiently find Pareto-optimal placements that balance wirelength, temperature, and warpage according to the design priority.*

III. ATMPlace Framework

This section presents the ATMPlace framework for thermo-mechanical-aware chiplet placement in 2.5D ICs. The overall architecture is introduced in Section III-A, followed by the compact thermal and warpage models in Sections III-B and III-C, respectively. The placement algorithm comprises initialization (Section III-D), multi-physics-aware optimization (Section III-E), and legalization (Section III-F).

A. Framework Overview

Figure 2 depicts the ATMPlace workflow. First, compact thermal and warpage models are trained offline using high-fidelity simulation data (Sections III-B–III-C). Subsequently, given a placement instance—defined by chiplet dimensions, microbump connectivity, and interposer specifications—the algorithm proceeds in three stages: (i) an MILP-based initialization generates a non-overlapping, orientation-aware seed layout (Section III-D); (ii) a gradient-based optimizer co-minimizes wirelength, thermal hotspots, and warpage using the compact models (Section III-E); (iii) a final MILP-based legalization stage eliminates residual overlaps and refines connectivity while preserving physical integrity (Section III-F).

B. Compact Thermal Model

Accurate and efficient computation of the temperature gradient ∇T with respect to chiplet locations is essential for gradient-based placement optimization. However, direct numerical simulation incurs prohibitive cost for large-scale systems [35]. Existing Green's function-based compact models [27, 28] discretize the domain into $M \times M$ grids and compute pairwise responses, yielding $\mathcal{O}(M^4)$ complexity for N chiplets. Moreover, these approaches model power sources as point or uniform-area heat sources and neglect material heterogeneity (e.g., silicon dies, epoxy underfill, Cu pillars), limiting accuracy for large-area, multi-material chiplets.

To overcome these limitations, we propose a chiplet-centric compact thermal model that: (i) treats each chiplet as a spatially extended, piecewise-homogeneous heat source; (ii) accounts for interfacial thermal resistance via effective length normalization; and (iii) reduces complexity to $\mathcal{O}(NM^2)$, where $N \ll M^2$ in practice.

We begin with the steady-state heat equation under quasi-homogeneous approximation:

$$\nabla^2 T(\mathbf{r}) = -\frac{P(\mathbf{r})}{\kappa(\mathbf{r})}, \tag{8}$$

where $T(\mathbf{r})$ is temperature, $P(\mathbf{r})$ is volumetric power density, and $\kappa(\mathbf{r})$ is thermal conductivity. Let $G(\mathbf{r}, \mathbf{r}')$ denote the free-space Green's function satisfying $\nabla^2 G(\mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}')$, with solution $G(\mathbf{r}, \mathbf{r}') = \frac{1}{4\pi\|\mathbf{r}-\mathbf{r}'\|}$. Applying Green's theorem yields the integral formulation:

$$T(\mathbf{r}) = -\int_\Omega G(\mathbf{r}, \mathbf{r}') \frac{P(\mathbf{r}')}{\kappa(\mathbf{r}')}\, d\mathbf{r}' = \sum_{i=1}^{N} \frac{P_i}{4\pi\kappa_i} \int_{\Omega_i} \frac{d\mathbf{r}'}{\|\mathbf{r} - \mathbf{r}'\|}, \tag{9}$$

where Ω_i is the 3D domain of chiplet i with uniform power density P_i and conductivity κ_i. The free-space assumption is justified by the interposer's lateral dimensions being significantly larger than individual chiplet footprints. To evaluate the volume integral efficiently, we apply the identity [36]:

$$\begin{aligned}\frac{1}{\|\mathbf{r} - \mathbf{r}'\|} &= \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-u^2\|\mathbf{r}-\mathbf{r}'\|^2}\, du \\ &= \frac{2}{\sqrt{\pi}} \int_0^\infty \iiint_{\Omega_i} e^{-u^2\left[(x-x')^2+(y-y')^2+(z-z')^2\right]} dx'\, dy'\, dz'\, du.\end{aligned} \tag{10}$$

Assuming chiplet thickness t_i is uniform and $t_i \ll w_i, h_i$, we approximate the z'-integral as $t_i e^{-u^2 a^2}$, where $a \in [0, t_i]$ is an effective decay depth (fitted during training). Utilizing the error function $\mathbf{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, the x' and y' integrals admit closed-form solutions $K(u, x, x_i, w_i) \triangleq$:

$$\frac{\sqrt{\pi}}{2u}\left[\mathbf{erf}\left(u\left(\frac{w_i'}{2} - (x - x_i)\right)\right) + \mathbf{erf}\left(u\left(\frac{w_i'}{2} + (x - x_i)\right)\right)\right]. \tag{11}$$

and similarly for y'. Substituting into (10) yields:

$$\frac{\sqrt{\pi} t_i}{2} \int_0^\infty \frac{e^{-u^2 a^2}}{u^2} K(u, x, x_i, w_i) K(u, y, y_i, h_i)\, du. \tag{12}$$

This integral reduces to a linear combination of four instances of the auxiliary function [36]:

$$F(a, b, c) = \int_0^\infty \frac{e^{-a^2 x^2}}{x^2} \operatorname{erf}(bx) \operatorname{erf}(cx)\, dx, \tag{13}$$

with closed-form expression:

$$F(a, b, c) = \frac{2}{\sqrt{\pi}} \left[b \ln \frac{c + \Delta}{\sqrt{a^2 + b^2}} + c \ln \frac{b + \Delta}{\sqrt{a^2 + c^2}} - a \tan^{-1} \frac{bc}{a\Delta} \right], \tag{14}$$

where $\Delta = \sqrt{a^2 + b^2 + c^2}$. To model heterogeneous thermal interfaces, we introduce anisotropic length normalization $l_{x,i}, l_{y,i} > 0$ per chiplet, yielding the final compact thermal model:

$$\begin{aligned} T_c(x, y) = \sum_{i=1}^{N} AP_i \Bigg[& F\left(a, \tfrac{w_i/2-(x-x_i)}{l_{x,i}}, \tfrac{h_i/2-(y-y_i)}{l_{y,i}}\right) \\ & + F\left(a, \tfrac{w_i/2-(x-x_i)}{l_{x,i}}, \tfrac{h_i/2+(y-y_i)}{l_{y,i}}\right) \\ & + F\left(a, \tfrac{w_i/2+(x-x_i)}{l_{x,i}}, \tfrac{h_i/2-(y-y_i)}{l_{y,i}}\right) \\ & + F\left(a, \tfrac{w_i/2+(x-x_i)}{l_{x,i}}, \tfrac{h_i/2+(y-y_i)}{l_{y,i}}\right) \Bigg] + B, \end{aligned} \tag{15}$$

where A (amplitude), a (decay depth), and B (bias) are global parameters. Thus, the model uses $2N+3$ trainable parameters.

Parameter fitting is performed offline: given a set of legal placements (typically 5–15, sufficient for convergence), ground-truth temperature fields $\{T_{\text{label}}\}$ are generated using numerical solver. The parameters $\boldsymbol{\beta} = \{A, a, B, \{l_{x,i}, l_{y,i}\}_{i=1}^N\}$ are optimized via: $\arg\min_{\boldsymbol{\beta}} \left\| T_c(\boldsymbol{\beta}) - T_{\text{label}} \right\|_2^2$ implemented in PyTorch. The closed-form structure of (15) ensures rapid convergence and robustness to initialization.

C. Compact Warpage Model

Full FEA-based warpage prediction is computationally infeasible for iterative placement. Inspired by the beam-theory analysis of Tsai and Wang [32], we derive a compact, differentiable surrogate model. Consider a three-layer package stack. The vertical displacement $D^z(r)$ along radial coordinate r (origin at package centroid) is:

$$D_1^z(r) = \frac{t\Delta\alpha\Delta T}{2\lambda D}\left(\tfrac{1}{2}r^2 - \tfrac{\cosh kr - 1}{k^2\cosh kl_1}\right), \quad 0 \le r \le l_1, \tag{16}$$

$$D_2^z(r) = D_1^z(l_1) + D_1^{z'}(l_1)(r-l_1) - \tfrac{1}{2}\nu_3 D_1^{z''}(0)(r-l_1)^2 \quad l_1 \le r \le l_2, \tag{17}$$

where t is total thickness, $\Delta\alpha$ is CTE difference, ΔT is thermal excursion, ν_i and E_i are Poisson's ratio and Young's modulus, $D_i = E_i t_i^3/[12(1-\nu_i)]$ is flexural rigidity, and $\lambda = \frac{1-\nu_1}{E_1 t_1} + \frac{1-\nu_3}{E_3 t_3} + \frac{t^2}{4D}$. Empirical studies indicate the second-order term in D_2^z contributes less than 5% to total warpage; thus, we approximate the displacement field as a localized quadratic form.

For each chiplet C_i, the warpage is modeled as a spatially varying quadratic–hyperbolic field centered at (x_i, y_i), where the anisotropy and effective decay in the x- and y-directions are captured by learnable scaling factors. Inspired by the analytical form in Tsai and Wang [32], but adapted for 2D in-plane placement and computational efficiency, we define the *local* warpage contribution of chiplet i as:

$$\begin{aligned} w_i(x,y) = &\left[k_{x,i}^2(x-x_i)^2 + k_{y,i}^2(y-y_i)^2\right] \\ &+ \lambda_i\left[k_{x,i}(x-x_i) + k_{y,i}(y-y_i)\right] + c_i, \end{aligned} \tag{18}$$

where $k_{x,i}, k_{y,i} > 0$ are anisotropic decay/curvature parameters (learnable), controlling the spatial *spread* of the warpage field—effectively normalizing physical distances by material-dependent thermal diffusion or mechanical stiffness scales; λ_i is a linear coupling coefficient, allowing asymmetry in the warpage gradient (e.g., due to boundary constraints or non-uniform underfill curing); and c_i is a constant offset representing baseline deformation associated with chiplet i.

The total warpage field is then obtained by thermally weighting each local contribution and summing:

$$W(x,y) = \alpha \cdot \sum_{i=1}^{N}\big(T(x,y) - T_{\text{ref}}\big)\cdot w_i(x,y) \;+\; b \tag{19}$$

Here, α is a global amplitude factor (shared across chiplets), capturing the overall thermo-mechanical coupling strength; T_{ref} is the reference temperature, accounting for process-induced pre-stress or ambient thermal history; and b is a global bias term, representing uniform out-of-plane offset (e.g., due to global CTE mismatch or post-molding shrinkage). Critically, the term $(T(x,y) - T_{\text{ref},i})$ takes the runtime local heating effects into account. Crucially, $W(x,y)$ is fully differentiable with respect to chiplet locations (x_i, y_i) and orientations (via T_c and w_i).

Model parameters $\boldsymbol{\gamma} = \{\alpha, b, \{k_{x,i}, k_{y,i}, \lambda_i, c_i, T_{\text{ref},i}\}_{i=1}^N\}$ (total $5N+2$) are fitted against FEA-generated warpage maps $\{W_{\text{label}}\}$: $\arg\min_{\boldsymbol{\gamma}} \left\|W(\boldsymbol{\gamma}) - W_{\text{label}}\right\|_2^2$. Convergence is accelerated by the physical priors embedded in (18)–(19): **Localized curvature** around each chiplet (via the quadratic term), proportional to local temperature rise; **Asymmetric tilting or shear** (via the linear term), arising from heterogeneous boundary conditions or placement-dependent stress concentration with only $5N+3$ learnable parameters (including α and b).

D. Initialization

A high-quality initial placement—jointly optimizing chiplet positions and orientations—is critical for convergence and solution quality [37]. To address this, we propose a MILP-based initializer that explicitly incorporates rotation and microbump clustering.

Let $\mathbb{A}$ denote the set of connected chiplet pairs. For each pair $(C_i, C_j) \in \mathbb{A}$, let A_{ij} be the number of D2D nets between them. Due to the structured layout of interface bumps (e.g., UCIe), these connections form spatially compact "clumps" rather than being uniformly distributed. We precompute the centroid offsets of each clump relative to its chiplet center: (O_{ij}^x, O_{ij}^y) on C_i and (O_{ji}^x, O_{ji}^y) on C_j, directly from the bump map. Each chiplet C_i supports four legal orientations: $\theta_i \in \{0°, 90°, 180°, 270°\}$. We encode θ_i using two binary variables $u_i, v_i \in \{0,1\}$ via the bijection:

$$\{0°, 90°, 180°, 270°\} \leftrightarrow \{(0,0),(0,1),(1,1),(1,0)\}.$$

The absolute position of the clump on C_i is then:

$$\begin{aligned} X_{ij} &= x_i + (1-u_i-v_i)O_{ij}^x - (v_i-u_i)O_{ij}^y, \\ Y_{ij} &= y_i + (v_i-u_i)O_{ij}^x + (1-u_i-v_i)O_{ij}^y, \end{aligned}$$

which yields linear expressions for all rotations—enabling exact MILP embedding. Using these, we define a bump-aware wirelength proxy for initialization:

$$\text{WL}^{(\text{init})} = \sum_{(i,j)\in\mathbb{A}} A_{ij}\left(|X_{ij} - X_{ji}| + |Y_{ij} - Y_{ji}|\right), \tag{20}$$

where the Manhattan distance approximates the total routed length of all A_{ij} nets between the two clumps.

Subject to the following constraints: (i) *Boundary containment*:

$$\frac{w_i'}{2} \le x_i \le W - \frac{w_i'}{2}, \quad \frac{h_i'}{2} \le y_i \le H - \frac{h_i'}{2}, \tag{21}$$

with rotated dimensions $w_i' = |1-u_i-v_i|\,w_i + |v_i-u_i|\,h_i$, $h_i' = |v_i-u_i|\,w_i + |1-u_i-v_i|\,h_i$. (ii) *Non-overlap*: enforced via four disjunctive constraints per chiplet pair (i,j) using auxiliary binaries $\delta_{ij}^{(k)} \in \{0,1\}$ $(k = 1,\ldots,4)$:

$$\begin{aligned} x_i + (w_i' + w_j')\varepsilon &\le x_j + W\delta_{ij}^{(1)}, \\ x_j + (w_i' + w_j')\varepsilon &\le x_i + W\delta_{ij}^{(2)}, \\ y_i + (h_i' + h_j')\varepsilon &\le y_j + H\delta_{ij}^{(3)}, \\ y_j + (h_i' + h_j')\varepsilon &\le y_i + H\delta_{ij}^{(4)}, \end{aligned} \tag{22}$$

where $\varepsilon \in [0, 0.5)$ is a small slack factor to ensure strict separation (e.g., $\varepsilon = 0.1$ yields 10% bin-width gap). Feasibility is guaranteed by:

$$\sum_{k=1}^{4}\delta_{ij}^{(k)} \le 3. \tag{23}$$

The full initialization problem is:

$$\min_{\mathbf{x},\mathbf{y},\mathbf{u},\mathbf{v},\boldsymbol{\delta}} \mathrm{WL}^{(\mathrm{init})} \quad \text{s.t.} \quad (21)\text{–}(23) \quad \text{holds},$$

solved efficiently using an off-the-shelf MILP solver. Though MILP scales exponentially in theory, our formulation remains tractable for dozens of chiplets—serving as a robust warm-start for subsequent gradient-based refinement.

E. Multi-Physics-Aware Optimization

Starting from the initialized layout, we perform gradient-based co-optimization of wirelength, density, temperature, and warpage. The objective is:

$$\mathcal{F}(\mathbf{X}) = \mathrm{WL}' + \lambda_{\text{dens}}\mathcal{D} + \lambda_T \|T_c(\cdot) - T_{\text{th}}\|_2^{\gamma} + \lambda_W \|W_c(\cdot) - W_{\text{th}}\|_2^{\gamma}, \tag{24}$$

where $\mathbf{X} = \{(x_i, y_i, \theta_i)\}_{i=1}^{N}$, WL' is the orientation-aware wirelength (see below), $\mathcal{D} = \sum_b (D'_b - M_b)^2$ penalizes bin density deviation from target M_b, $T_c(\cdot)$ and $W_c(\cdot)$ are the outputs of the compact thermal and warpage models, and they are differentiable w.r.t. x_i, y_i and θ_i through the chain rule.

1) Orientation-aware wirelength and density: Our orientation-aware optimization algorithm is based on the model proposed by Lin et al. [34]. Chiplets admit discrete rotations $\Theta = \{0°, 90°, 180°, 270°\}$. We use a smooth projection that calculates the probability of the chiplet C_i to be rotated to each legal orientation θ^k [34]:

$$B_z(\theta^k, \theta_i) = \frac{\exp\left(R_z(\theta^k, \theta_i)/\eta\right)}{\sum_{\theta^k \in \Theta} \exp\left(R_z(\theta^k, \theta_i)/\eta\right)}, \tag{25}$$

where $\eta > 0$ controls sharpness. With angular deviation $\Delta\theta_i^k$, $R_z(\theta^k, \theta_i)$ defined piecewise as:

$$\begin{aligned} \Delta\theta_i^k &= \left|0.5 - \left|0.5 - \|\theta_i - \theta^k\|/360\right|\right|, \\ R_z\left(\theta^k, \theta_i\right) &= \begin{cases} 1 - 2|\Theta|^2 \left|\Delta\theta_i^k\right|^2, & 0 < \Delta\theta_i^k \le \frac{1}{2|\Theta|} \\ 2|\Theta|^2 \left(\Delta\theta_i^k - \frac{1}{|\Theta|}\right)^2, & \frac{1}{2|\Theta|} < \Delta\theta_i^k \le \frac{1}{|\Theta|} \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{26}$$

The projected wirelength and density become:

$$\mathrm{WL}' = \sum_{e\in\mathbb{E}} \sum_{p_i,p_j\in\mathcal{B}_e} \sum_{\theta^k,\theta^\ell\in\Theta} B_z(\theta^k,\theta_i) B_z(\theta^\ell,\theta_j)\, \mathrm{HPWL}(p_i^{(k)}, p_j^{(\ell)}), \tag{27}$$

$$D'_b = \sum_{C_i\in\mathbb{C}} \sum_{\theta^k\in\Theta} B_z(\theta^k,\theta_i) \cdot \mathrm{Area}(C_i^{(k)} \cap \text{bin } b). \tag{28}$$

2) Optimization algorithm: In our framework, the minimization of (7) is solved by the conjugate gradient descent (CGD) method. The algorithm is summarized in Algorithm 1 with adaptive step sizing and noise injection to escape local minima. Step sizes for position and angle are decoupled: $\alpha_k = \left(\alpha_{k,x}, \alpha_{k,y}, \alpha_{k,\theta}\right)$. The density weight λ_{dens} is adaptively scaled:

$$\lambda_{\text{dens}}^{(0)} = \frac{\sum|\nabla_{\mathbf{x},\mathbf{y}}\mathrm{WL}'|}{\sum|\nabla_{\mathbf{x},\mathbf{y}}\mathcal{D}|}, \quad \lambda_{\text{dens}}^{(k+1)} = \lambda_{\text{dens}}^{(k)} \cdot \left(1 + \rho \cdot \mathrm{OVFL}^{(k)}\right). \tag{29}$$

Finally, the system is prone to get stuck in local optima during the nonlinear optimization. Inspired by [38], we perturb the placement by adding random noise when the overflow of the system is stuck at high values or converged. The overflow metric reads:

$$\mathrm{OVFL} = \frac{1}{S}\sum_b \max\left(0, D'_b - M_b\right), \tag{30}$$

where S is the total chiplet area.

Algorithm 1 Optimization via CGD

Input: $\mathcal{F}(\mathbf{X})$: objective function; Max_iter: number of max iteration; $\mathbf{lr} = (lr_{pos}, lr_{ang})$: learning rate of position and angular variables

Output: optimal $\mathbf{X}^*$

1: initialize $\mathbf{X}_0$ (Section III-D), λ_{dens}, $\mathbf{g}_0 = \mathbf{0}$, and $\mathbf{d}_0 = \mathbf{0}$;
2: **for** $k = 1$ to Max_iter **do**
3: Compute $\mathcal{F}(\mathbf{X}_{k-1})$ and gradient $\mathbf{g}_k \leftarrow \nabla\mathcal{F}(\mathbf{X}_{k-1})$
4: $\beta_k \leftarrow \frac{\mathbf{g}_k^\top(\mathbf{g}_k - \mathbf{g}_{k-1})}{\|\mathbf{g}_{k-1}\|_2^2 + \epsilon}$ (Polak–Ribière, $\beta_1 = 0$)
5: $\mathbf{d}_k \leftarrow -\mathbf{g}_k + \beta_k \mathbf{d}_{k-1}$
6: $\alpha_k \leftarrow \left(\eta_x/\|\mathbf{d}_k^{(x)}\|,\ \eta_y/\|\mathbf{d}_k^{(y)}\|,\ \eta_\theta/\|\mathbf{d}_k^{(\theta)}\|\right)$
7: $\mathbf{X}_k \leftarrow \mathbf{X}_{k-1} + \alpha_k \odot \mathbf{d}_k$
8: legalize (Section III-F) and derive the final result $\mathbf{X}^*$

F. Legalization

After optimization, the chiplets are roughly uniformly distributed across the interposer. But there are still some overlapping areas at this time. Additionally, sometimes minor temperature optimization can lead to excessive increases in total wirelength. Thus, a legalization stage is indispensable to eliminate all overlaps and reduce the total wirelength while keeping the changes in chiplet positions minimal, essentially maintaining the temperature distribution. In this stage, given the optimized solution $\left(x_i^{(opt)}, y_i^{(opt)}, \theta_i^{(opt)}\right)$, we fix the orientations and optimize the positions of chiplets by MILP.

The optimization objective includes two parts. The first term seeks to minimize the total wirelength:

$$\begin{aligned} WL^{(legal)} = \Sigma_{e\in\mathbb{E}}\Sigma_{p_i,p_j\in\mathbb{B}_e} \Big(&\|(x_i + x_{p_i}) - (x_j + x_{p_j})\| \\ &+ \|(y_i + y_{p_i}) - (y_j + y_{p_j})\| \Big), \end{aligned} \tag{31}$$

While the second term constrains the displacement:

$$DSP = \Sigma_{C_i}\left(\|x_i - x_i^{(opt)}\| + \|y_i - y_i^{(opt)}\|\right). \tag{32}$$

Combined with the non-overlapping conditions formulated before ((22) with $\varepsilon = 1$) with the chiplet width and height modified to involve the minimum spacing between chiplets, the final legalization problem is:

$$\min_{\mathbf{x},\mathbf{y}} DSP + \lambda_w \cdot WL^{(legal)} \quad \text{s.t.} \quad (21)\text{–}(23) \quad \text{holds},$$

Here λ_w is a parameter that controls the extent to which the total wirelength is optimized in this stage. Since it may be time-consuming or sometimes even infeasible to optimize the wirelength, the algorithm will set λ_w to 0 when the time limit of 10 minutes reached.

Ending on a note, experiments have shown that the quality of the placement results heavily relies on the choice of parameters. Thus, inspired by AutoDMP [37], which puts forth to optimize the hyperparameters through Bayesian optimization to improve overall performance, we also carry out hyperparameter optimization to find the most favorable parameters.

IV. Experimental Results

This section presents a comprehensive evaluation of the proposed ATMPlace framework. We first describe the experimental setup in Section IV-A, followed by validation of the compact thermal and warpage models. A detailed analysis of placement quality, runtime, and multi-objective trade-offs is provided in Section IV-C.

A. Experimental Setup

Given the scarcity of publicly available large-scale 2.5D-IC benchmarks and the omission of die-to-die (D2D) interconnects in prior works, we construct ten representative test cases, summarized in Table III. These cases span 6 to 61 chiplets, up to thousands of nets, and whitespace ratios between 0.35 and 0.65. The first five cases are high-performance computing (HPC) systems comprising CPUs, GPUs, HBMs, and DRAMs, adapted from [3]. The latter five incorporate analog and micro-electromechanical system (MEMS) modules [39, 40] to evaluate heterogeneous integration capability. Chiplet power densities range from $2 \times 10^5\,\mathrm{W/m^2}$ to $3 \times 10^6\,\mathrm{W/m^2}$.

For thermal simulation, we use **ATSim** [24] as the golden reference due to its capability in modeling multiple thermal properties, like anisotropy and heterogeneity. Mechanical warpage is simulated using an in-house finite-element solver [41]. Both simulators are verified and calibrated against the ANSYS commercial simulators. Material parameters and boundary conditions follow established conventions in the literature [41, 42]. A uniform grid resolution of 64×64 is employed.

Inter-chiplet communication is enabled via D2D link modules. We adopt both the *standard* ($\times$16) and *advanced* ($\times$32) UCIe-compliant interfaces. The $\times$16 module offers longer channel reach, while the $\times$32 variant provides higher bandwidth. Key interface parameters are listed in Table IV, with bump pitch geometries illustrated in Fig. 3.

The ATMPlace framework[2] is implemented in Python, building upon DREAMPlace [43] for infrastructure and Optuna [44] for hyperparameter optimization. MILP subproblems are solved using Gurobi [45].

All experiments are conducted on a Linux server equipped with dual Intel Xeon Silver 4210 CPUs (2.20 GHz, 20 cores total), 128 GB RAM, and a maximum time budget of 12 hours per run. For fairness, all baseline methods—including enumeration-based SP-CP—are allowed up to 80 CPU cores and the same time limit.

[2]Benchmarks and binaries are open-sourced at https://github.com/Brilight/ATPlace_pub.

TABLE III: Benchmark configurations.

Case	Bump Type	Dies	Nets	Interposer		
				Width (mm)	Height (mm)	Whitespace/%
1	$\times$32	6	3168	42.0	42.0	40
2	$\times$32	6	3520	55.0	52.0	65
3	$\times$32	8	8448	39.0	39.0	60
4	$\times$32	11	7040	57.0	59.0	40
5	$\times$32	12	7392	37.0	37.0	35
6	$\times$16	20	5632	49.0	53.0	55
7	$\times$16	28	2816	30.0	25.0	55
8	$\times$16	36	2948	26.0	23.0	60
9	$\times$16	44	7656	59.0	55.0	45
10	$\times$16	61	5280	47.0	47.0	50

TABLE IV: Characteristics of UCIe-compliant interfaces.

Package	Cols	Lanes	Bump Pitch (μm)	Pitch$_x$ (μm)	Pitch$_y$ (μm)
Standard ($\times$16)	12	16	100	180	90
Advanced ($\times$32)	16	32	25	27	42

B. Compact Models

We first assess the fidelity of the proposed compact models, which serve as the foundation for gradient-based placement optimization. For each benchmark case, we generate 20 random placements for training and testing. Model accuracy is quantified via mean absolute error (MAE) and Pearson correlation coefficient (ρ). While MAE measures absolute deviation, ρ evaluates not only global physical field alignment but also gradient similarity—critical for optimization stability and convergence.

As shown in Table V, the compact thermal model achieves an average MAE of $2.4\,^\circ\mathrm{C}$ and $\rho = 0.977$, indicating excellent agreement with ATSim. Compared to the golden simulator (average runtime 66 s per layout), our model reduces inference time to 7.5 ms on average—a speedup exceeding $8000\times$. Similarly, the compact warpage model attains an average MAE of $2.5\,\mu\mathrm{m}$ and $\rho = 0.978$, with a $> 10000\times$ speedup over the numerical mechanical simulator.

Figure 4 visualizes the temperature and displacement fields for Case 1. The predicted temperature distributions closely match the ground truth, with errors primarily localized at: (1) chiplet edges—attributable to the quasi-homogeneous approximation in (8); and (2) interposer boundaries—stemming from the infinite-space Green's function assumption. The latter could be mitigated via image-source methods (e.g., [46]), albeit at increased computational cost.

For the displacement field, the dominant source of prediction error arises from the modeling simplification of the linear dependence on local temperature fields (19), which neglects higher-order thermo-mechanical coupling. In reality, the displacement field tends to be spatially smoother than the raw temperature field due to structural continuity—our current linear surrogate, while efficient, cannot fully capture this regularization effect.

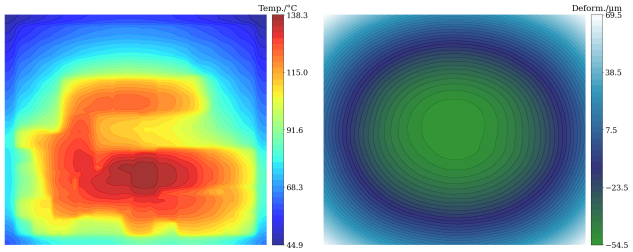
C. Placement Results

We evaluate four representative placement methodologies:

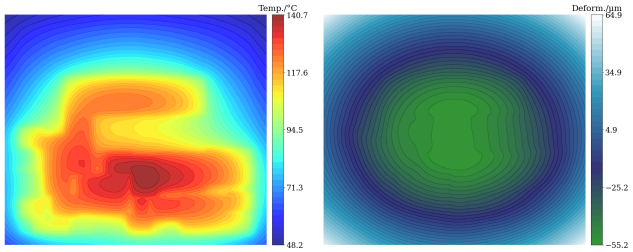
- TAP-2.5D [3]: Simulated Annealing (SA)-based, open-sourced and adapted to our benchmark;
- TACPlace [4]: GA/PSO hybrid, representing evolutionary approaches, realized and adapted for thermo-mechanical applications;

TABLE V: Performance of the proposed compact models. RT denotes runtime; Corr. denotes the correlation coefficient.

Case	Thermal				Mechanical			
	Golden	Ours			Golden	Ours		
	RT/s	RT/ms	MAE/°C	Corr.	RT/s	RT/ms	MAE/μm	Corr.
1	70	4.7	2.15	0.984	5	0.3	2.03	0.989
2	67	4.9	0.97	0.991	7	0.3	2.70	0.979
3	82	6.7	3.13	0.984	5	1.1	2.17	0.973
4	78	5.9	3.18	0.977	8	0.5	3.16	0.989
5	87	6.3	3.95	0.982	5	0.5	2.27	0.980
6	53	10.2	2.70	0.965	8	0.7	2.77	0.979
7	58	11.4	1.92	0.955	6	0.6	1.19	0.969
8	56	19.1	2.18	0.960	6	1.2	0.65	0.986
9	49	16.0	2.02	0.992	13	0.7	4.07	0.985
10	60	22.1	2.13	0.981	12	0.8	4.14	0.954
Avg.	**8680×**	**1×**	**2.43**	**0.977**	**13050×**	**1×**	**2.52**	**0.978**



(a) Golden truth



(b) Predictions

Fig. 4: Thermo-mechanical field comparison for a placement in Case 1: (a) ground truth, (b) compact model prediction.

- `SP-CP` [7]: enumeration-based (times out for cases with more than 12 chiplets), wirelength-driven binary tested only;
- `ATPlace2.5D` [1]: our prior work, thermal-aware only;
- `ATMPlace`: the proposed framework, jointly optimizing thermal and mechanical warpage.

Reinforcement learning (RL)-based methods [8, 9] are excluded due to unavailability of executables; reported results suggest their performance is comparable to or slightly better than `TAP-2.5D`. The method of [6] fails on our benchmarks due to scalability limitations.

1) Wirelength-Driven Optimization: For completeness and to isolate the impact of physical-aware objectives, we first evaluate pure wirelength-driven (WL-driven) optimization. As shown in Table VI, `ATMPlace` achieves state-of-the-art wirelength minimization: it delivers the shortest TWL in 8 out of 10 cases, and ties `SP-CP` in Case 1 and 3. The average TWL is 1.00× (baseline), compared to 3.0× and 3.9× for `TAP-2.5D` and `TACPlace`, respectively. Even in Case 5—the largest enumerated instance (`SP-CP` times out with $< 52.052\,\mathrm{m}$ TWL)—`ATMPlace` obtains TWL $= 48.265\,\mathrm{m}$, which is $> 2.4\times$ shorter than `TAP-2.5D` and $> 2.3\times$ shorter than `TACPlace`. Moreover, the runtime advantage of `ATMPlace` persists in WL mode: with average runtime $= 2.3\,\mathrm{min}$, it is $14.2\times$ faster than `SP-CP` (on solvable cases) and $> 3\times$ faster than heuristic methods. The enumeration approach, while yielding marginally better TWL (e.g., in Case 1), fails to scale beyond 12 chiplets and produces thermally hazardous layouts.

This superiority arises from the synergy of three components: (i) the MILP-based initialization, which yields a high-quality seed layout with orientation-aware bump alignment; (ii) the smooth orientation parametrization, which enables gradient-based refinement of discrete rotations without combinatorial explosion; and (iii) the adaptive density-weight scheduling (30), which dynamically balances HPWL reduction against placement overflow—preventing premature congestion that traps heuristic methods.

However, WL-driven optimization incurs significant physical penalties. For instance, in Case 3, `ATMPlace` achieves TWL $= 32.350\,\mathrm{m}$ (vs. $134.073\,\mathrm{m}$ in TM mode) but suffers $T_{\max} = 143.8\,^{\circ}\mathrm{C}$—an increase of $36.5\,^{\circ}\mathrm{C}$ over the TM-aware variant—and warpage rises from $34.1\,\mu\mathrm{m}$ to $66.7\,\mu\mathrm{m}$. Similar trends hold universally: minimizing TWL alone encourages tight clustering of connected chiplets (e.g., GPUs with adjacent HBMs), intensifying local hotspots and bending moments. This confirms the *necessity* of multi-physics co-optimization: wirelength-driven placement, while optimal in isolation, is physically infeasible for advanced 2.5D-ICs under stringent thermo-mechanical reliability constraints.

2) Thermo-Mechanical-Aware Optimization: We summarize the thermo-mechanical-aware placement results in Table VII. Across all ten benchmarks, `ATMPlace` consistently achieves the best trade-off among wirelength, thermal, and mechanical objectives on average. The average gains are decisive: normalized to `ATMPlace`, `TAP-2.5D` and `TACPlace` incur 2.5× and 1.5× longer TWL, 3%–13% higher $T_{\max}$, and 5%–27% larger warpage—at 7×–12× runtime overhead. Runtime averages 31 min—7.5× faster than `TAP-2.5D`—with per-run optimization costing only 20–40% of total time after one-time model training (Fig. 7).

Besides, `ATMPlace` yields the lowest $T_{\max}$ in 6 cases, the shortest TWL in 5 cases, and the smallest warpage in 5 cases. Notably, it simultaneously minimizes all three metrics in Case 5 and 6. In Case 5, `ATMPlace` reduces $T_{\max}$ by $19.1\,^{\circ}\mathrm{C}$ and warpage by $6.3\,\mu\mathrm{m}$ over `TAP-2.5D`, while shortening TWL by 40%. Even against our thermal-optimization-only predecessor `ATPlace2.5D`, `ATMPlace` lowers warpage by up to 60% (Case 3) without degrading $T_{\max}$. Thus, `ATMPlace` is the only method achieving scalable, high-quality TM co-optimization. This stems from three key features: (i) *Joint differentiable modeling* of thermal and warpage fields, enabling gradient-guided symmetry enforcement (e.g., mirrored placement of high-power dies); (ii) *Warpage-aware regularization* via (19), which suppresses torque-inducing layouts missed by thermal-only optimization; (iii) *Decoupled step-size control* for position/orientation (Algorithm 1), preventing angular updates from destabilizing mechanical equilibrium.

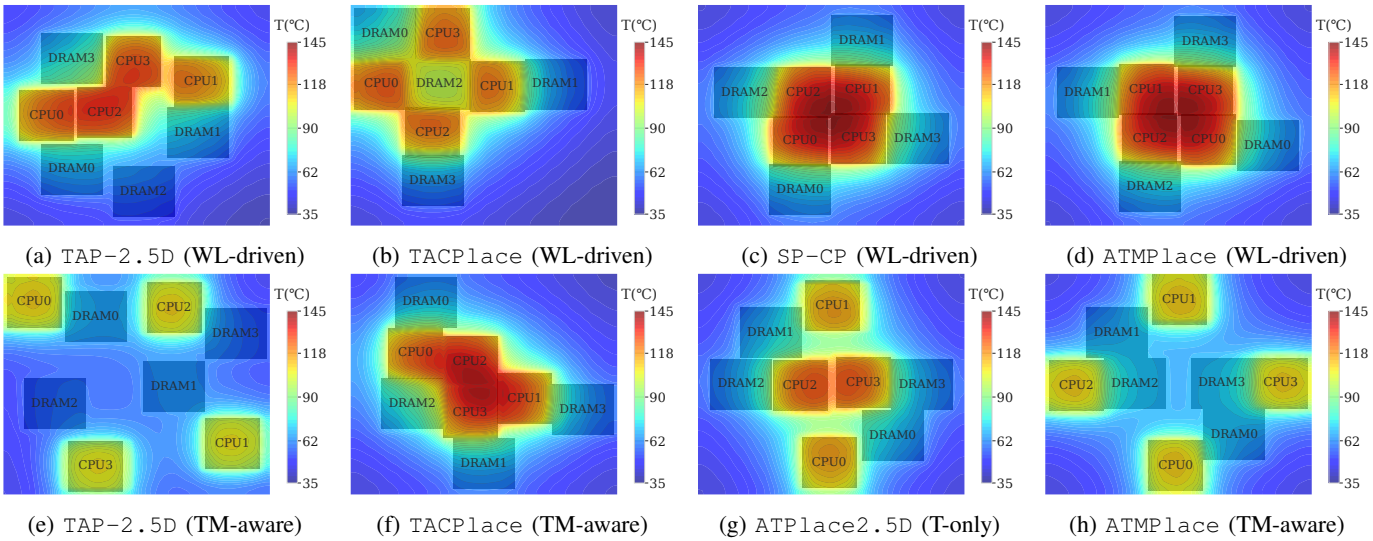
TABLE VI: Comparison of placement algorithms under wirelength-driven optimization. 'Wpg' represents the warpage. Normalized metrics use `ATMPlace` as baseline (1×).

Case	TAP-2.5D [3]				TACPlace [4]				SP-CP [7]				ATMPlace (WL-driven)			
	RT/min	TWL/m	T_{max}/°C	Wpg/μm	RT/min	TWL/m	T_{max}/°C	Wpg/μm	RT/min	TWL/m	T_{max}/°C	Wpg/μm	RT/min	TWL/m	T_{max}/°C	Wpg/μm
1	3.1	27.499	101.5	84.6	2.0	60.537	100.2	86.4	0.02	**11.902**	101.1	84.9	0.3	11.973	102.7	85.1
2	3.5	62.124	66.4	97.4	2.1	74.656	66.0	97.6	0.01	16.107	67.4	103.8	0.3	**15.408**	67.8	105.3
3	7.3	87.543	132.8	63.7	5.5	122.972	126.1	60.9	0.5	**32.200**	144.2	66.7	0.7	32.350	143.8	66.7
4	6.3	150.766	106.4	131.6	5.0	187.040	113.4	143.1	84	54.306	114.8	147.4	1.5	**46.982**	114.8	143.9
5	6.3	119.159	147.1	71.3	5.2	112.249	161.2	76.6	> 12 h	< 52.052†	N/A	N/A	2.4	**48.265**	162.2	76.6
6	5.0	68.657	99.7	92.0	4.2	125.205	105.5	112.7	Timeout				1.1	**29.482**	111.3	104.6
7	2.8	29.094	82.5	29.1	3.1	34.213	83.3	32.6	Timeout				2.7	**12.523**	86.0	33.1
8	3.0	20.065	79.3	22.0	3.5	30.611	81.2	26.0	Timeout				3.2	**9.001**	82.4	23.0
9	7.0	174.850	105.5	130.3	4.8	207.050	114.6	142.9	Timeout				4.1	**43.401**	124.3	145.2
10	4.9	121.066	104.1	84.5	6.2	97.650	111.3	94.6	Timeout				7.5	**26.912**	125.5	91.5
Avg.	4.813×	3.012×	0.922×	0.926×	3.508×	3.879×	0.952×	1.006×	14.204×	1.048×	0.995×	1.002×	**1×**	**1×**	**1×**	**1×**

† The binary of `SP-CP` only reports the temporary minimal total wirelength achieved at the time limit

TABLE VII: Comparison under thermo-mechanical-aware optimization. Normalized metrics use `ATPlace2.5D` as baseline (1×) for runtime and warpage, and `ATMPlace` for wirelength and temperature.

Case	TAP-2.5D [3]				TACPlace [4]				ATPlace2.5D [1]				ATMPlace (TM-aware)			
	RT/h	TWL/m	T_{max}/°C	Wpg/μm	RT/h	TWL/m	T_{max}/°C	Wpg/μm	RT/h	TWL/m	T_{max}/°C	Wpg/μm	RT/h	TWL/m	T_{max}/°C	Wpg/μm
1	3.3	89.887	95.3	77.1	5.3	62.040	100.6	85.6	0.3	34.560	92.2	77.2	0.4	25.900	94.6	76.5
2	3.9	163.200	65.5	71.8	5.0	97.773	66.5	95.2	0.3	86.901	64.1	83.0	0.3	56.807	64.9	84.3
3	3.1	254.694	110.0	51.5	4.5	114.597	139.7	69.3	0.3	132.964	111.3	55.3	0.4	219.062	107.3	34.1
4	4.2	364.663	108.9	125.4	4.8	180.313	111.4	147.0	0.3	144.987	99.1	126.2	0.4	150.416	100.6	128.3
5	3.7	198.420	142.3	66.0	6.5	116.413	163.8	73.7	0.4	159.714	130.7	64.3	0.5	117.776	123.2	60.0
6	3.2	196.987	89.8	89.8	6.8	121.524	102.2	105.3	0.3	99.739	86.5	83.0	0.5	82.236	83.3	79.3
7	2.1	52.370	78.8	27.4	6.8	33.602	82.7	32.6	0.5	35.662	71.2	27.6	0.5	36.119	70.7	27.6
8	2.3	49.711	77.1	17.8	7.4	32.767	79.5	24.2	0.5	29.369	69.7	15.2	0.6	23.680	69.8	16.4
9	5.6	328.317	106.0	127.4	8.4	194.374	118.5	144.0	0.8	62.277	128.5	141.5	0.9	70.717	118.6	140.0
10	5.2	144.716	101.9	82.7	8.4	106.826	120.1	97.1	1.1	53.336	119.8	88.7	1.2	59.547	125.2	90.8
Avg.	7.189 ×	2.312 ×	1.028 ×	1.048 ×	12.203 ×	1.516 ×	1.132 ×	1.274 ×	0.829 ×	1.101 ×	1.013 ×	1.063 ×	1×	**1×**	**1×**	**1×**



(a) `TAP-2.5D` (WL-driven) (b) `TACPlace` (WL-driven) (c) `SP-CP` (WL-driven) (d) `ATMPlace` (WL-driven)

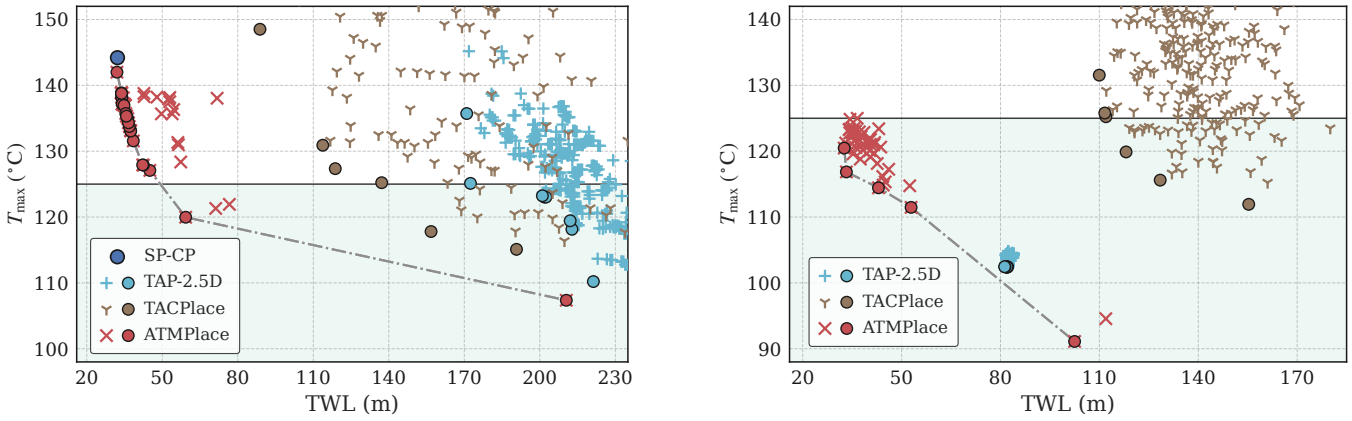
(e) `TAP-2.5D` (TM-aware) (f) `TACPlace` (TM-aware) (g) `ATPlace2.5D` (T-only) (h) `ATMPlace` (TM-aware)

Fig. 5: Placement layouts for Case 3. Upper row: wirelength-driven optimization. Lower row: thermo-mechanical-aware optimization. CPU and DRAM dimensions: 9×8 mm^2 and 9×9 mm^2, respectively.

3) Placement Layouts: Figure 5 illustrates the placement outcomes for Case 3 under wirelength-driven (WL-driven) and thermo-mechanical-aware (TM-aware) optimizations, revealing distinct layout strategies dictated by objective formulation.

Under wirelength-driven optimization, the enumeration-based `SP-CP` (Fig. 5(c)) achieves the global WL minimum by aggregating the four CPUs at the center and arranging DRAMs radially—a layout that minimizes interconnect distances but concentrates thermal power density. Remarkably, `ATMPlace` (Fig. 5(d)) reproduces this near-optimal topology with only 0.5% WL overhead, demonstrating the efficacy of its gradient-based refinement from a high-quality initial solution. In contrast, heuristic methods exhibit structural inefficiencies: `TAP-2.5D` (Fig. 5(a)) splits the CPUs, increasing inter-CPU and CPU–DRAM wirelength, while `TACPlace` (Fig. 5(b)) yields an elongated, loosely packed configuration—both indicative of premature convergence in high-dimensional discrete spaces.

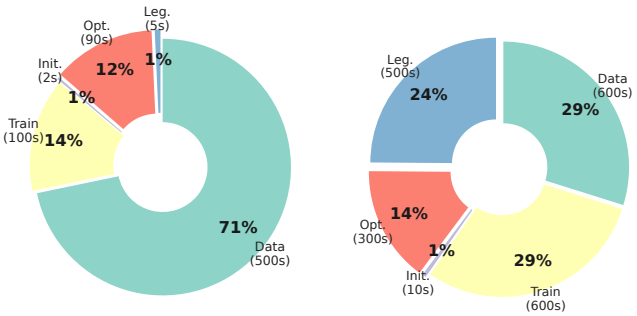
In thermo-mechanical-aware optimization, physical constraints dominate layout morphology. `TAP-2.5D` (Fig. 5(e)) adopts conservative clustering, resulting in low warpage (52 μm) but poor thermal performance ($T_{\max} = 110\,^{\circ}$C) and excessive wirelength (+124% over `ATMPlace`). `TACPlace`



(a) Case 3 (b) Case 10

Fig. 6: Multi-objective trade-offs between maximum temperature ($T_{\max}$) and total wirelength (TWL) for Case 3 (left) and Case 10 (right). Solid circles represent the Pareto-optimal solutions obtained by each method. The dashed line indicates an approximate Pareto frontier constructed from the best-known solutions across all methods.

(Fig. 5(f)) partially disperses CPUs yet retains asymmetry, leading to the highest temperature and warpage (140 °C, 69 µm) among TM-aware methods. The thermal-only `ATPlace2.5D` (Fig. 5(g)) improves temperature ($T_{\max}$ = 111 °C) via CPU separation but overlooks mechanical coupling—its staggered DRAM placement induces torque-like bending, reflected in elevated warpage (55 µm). Only `ATMPlace` (Fig. 5(h)) achieves full co-optimization: CPUs are symmetrically positioned, DRAMs are mirrored across both axes, and inter-chiplet spacing is uniformly enlarged in high-gradient zones. This layout delivers balanced $T_{\max}$ = 107 °C and Wpg = 34 µm—matching the thermal performance of `ATPlace2.5D` while ensuring mechanical robustness through spatial symmetry enforced by the warpage-aware gradient term.



(a) Case 3 (b) Case 10

Fig. 7: Runtime breakdown of `ATMPlace` in thermo-mechanical-aware optimization. 'Data.': dataset generation; 'Train.': compact model training; 'Init.': initialization; 'Opt.': optimization; 'Leg.': legalization. Percentages indicate the share of the total runtime.

4) Multi-Objective Optimization: Figure 6 quantifies the trade-offs between $T_{\max}$ and TWL for Case 3 (left) and Case 10 (right). `ATMPlace` generates a dense, continuous

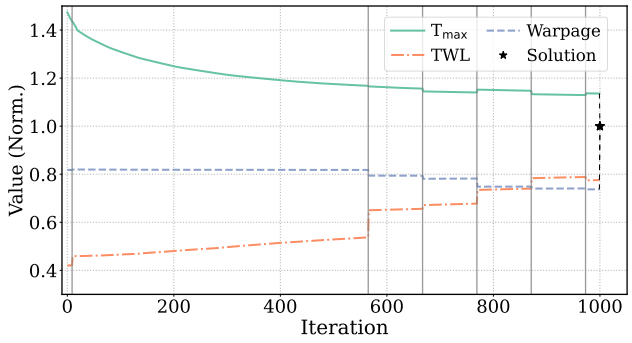


Fig. 8: Normalized convergence trajectories of $T_{\max}$, wirelength (TWL), and warpage under thermo-mechanical-aware optimization (Case 3). Curves show evolution before the final solution (marked by ⋆); the vertical lines denote entropy-driven noise injections. All metrics are normalized to their final values.

Pareto front spanning from wirelength-optimal to thermo-mechanical-optimal, demonstrating full control over the design space.

In Case 3, `ATMPlace` dominates all baselines: its solutions form a tight, low-$T_{\max}$ cluster below 125 °C, while `TAP-2.5D` and `TACPlace` are confined to higher wirelength region. The enumeration-based `SP-CP` achieves minimal TWL but suffers extreme thermal penalty ($T_{\max}$ > 140 °C). For the large-scale Case 10, `ATMPlace` again establishes the frontier, while `TAP-2.5D` clusters, and `TACPlace` fail to enter the optimal region. This highlights `ATMPlace`'s scalability and ability to navigate complex multi-objective landscapes where heuristic methods stagnate. To conclude, `ATMPlace` establishes the Pareto frontier: it dominates all baselines in the three-dimensional objective space of (TWL, temperature, warpage).

5) Runtime Analysis: Figure 7 details the runtime breakdown for Case 3 and Case 10. Model construction (dataset

generation + training) dominates ($85\,\%$ for Case 3, $58\,\%$ for Case 10), but is a one-time cost amortized over multiple runs (e.g., Pareto exploration). Initialization is negligible ($< 10\,\mathrm{s}$). The core optimization and legalization stages consume $10\,\%$–$40\,\%$ of total time, reflecting their computational complexity.

Figure 8 shows the normalized convergence trajectories for Case 3 under thermo-mechanical-aware optimization. As high-power chiplets disperse to mitigate thermal hotspots, $T_{\max}$ drops sharply in early iterations (from $1.4\times$ to $\sim 1.15\times$ its final value), while TWL initially rises—reflecting the trade-off between thermal relief and interconnect cost—before stabilizing near its final value. Warpage exhibits a gradual decline, indicating progressive mechanical stress reduction throughout the optimization. Periodic entropy-driven noise injections (dashed vertical lines) perturb the layout when the optimization gets stuck, enabling escape from local minima and yielding incremental improvements across all three objectives. Notably, after each injection, the system converges faster to a lower-cost state, demonstrating that the perturbation enhances exploration without destabilizing the solution trajectory.

The robustness of our approach is further evidenced by the smooth, deterministic convergence profile: unlike stochastic methods (e.g., SA, GA), where solution quality varies significantly with random seeds—posing a reliability risk in the design flow—our method consistently reaches the final solution ($\star$). The final solution (at iteration 1000) represents a balanced Pareto-optimal trade-off among thermal, mechanical, and wiring constraints.

V. Conclusion

As the semiconductor industry increasingly embraces 2.5D-ICs to overcome the limitations of traditional scaling, the design of large-scale, heterogeneous chiplet-based systems has become both urgent and inevitable. This paper presents `ATMPlace`, a scalable analytical placement framework for large-scale 2.5D-ICs that jointly optimizes wirelength, thermal integrity, and mechanical reliability—addressing a critical gap in existing methodologies. Our physics-based compact thermal-mechanical model attains a mean correlation above 0.97 and accelerates simulation by over $8000\times$ compared to numerical solvers, validating its fidelity for gradient-based optimization. By integrating orientation-aware analytical placement with fast compact models, `ATMPlace` achieves high-quality layouts at unprecedented speed: it delivers up to 146% and 52% better total wirelength, 3% and 13% lower peak temperature, and 5% and 27% better warpage than `TAP-2.5D` and `TACPlace`, respectively, under thermo-mechanical-aware constraints. In wirelength-driven mode, it matches the near-optimal `SP-CP` within 1% deviation.

Looking ahead, we plan to extend the framework to support co-design with interposer routing by incorporating congestion and signal integrity into the objective, enabling joint placement-routing optimization. We believe `ATMPlace` will lay a critical foundation for scalable, reliable, and automated design of next-generation 2.5D-IC systems.

VI. Acknowledgment

The authors thank Xueqing Li, Yifan Chen, Xinming Wei for insightful discussions. This paper used GPT-5 to generate the thermal deformation schematic figure in Sec. I.

References

[1] Q. Wang, X. Li, T. Jia, Y. Lin, R. Wang, and R. Huang, "Atplace2. 5d: Analytical thermal-aware chiplet placement framework for large-scale 2.5 d-ic," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024, pp. 1–9.

[2] Y.-K. Ho and Y.-W. Chang, "Multiple chip planning for chip-interposer codesign," in *Proceedings of the 50th Annual Design Automation Conference*, 2013, pp. 1–6.

[3] Y. Ma, L. Delshadtehrani, C. Demirkiran, J. L. Abellan, and A. Joshi, "Tap-2.5 d: A thermally-aware chiplet placement methodology for 2.5 d systems," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 1246–1251.

[4] S. Yu, H. Liu, X. Wei, B. Shi, and G. Luo, "Tacplace: Ultrafast thermal-aware chiplet placement with feasibility seeking," in *Proceedings of the Great Lakes Symposium on VLSI 2025*, ser. GLSVLSI '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 600–605.

[5] P. Zhang, D.-W. Wang, and W.-S. Zhao, "A thermal and power integrity co-optimization framework for 2.5-d integrated microsystem," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 3, pp. 1397–1410, 2025.

[6] S. Osmolovskyi, J. Knechtel, I. L. Markov, and J. Lienig, "Optimal die placement for interposer-based 3d ics," in *2018 23rd Asia and South Pacific design automation conference (ASP-DAC)*. IEEE, 2018, pp. 513–520.

[7] H.-W. Chiou, J.-H. Jiang, Y.-T. Chang, Y.-M. Lee, and C.-W. Pan, "Chiplet placement for 2.5 d ic with sequence pair based tree and thermal consideration," in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, 2023, pp. 7–12.

[8] Y. Duan, X. Liu, Z. Yu, H. Wu, L. Shao, and X. Zhu, "Rlplanner: Reinforcement learning based floorplanning for chiplets with fast thermal analysis," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2024.

[9] Z. Deng, Y. Duan, L. Shao, and X. Zhu, "Chiplet placement order exploration based on learning to rank with graph representation," *arXiv preprint arXiv:2404.04943*, 2024.

[10] S. Chen, S. Li, Z. Zhuang, S. Zheng, Z. Liang, T.-Y. Ho, B. Yu, and A. L. Sangiovanni-Vincentelli, "Floorplet: Performance-aware floorplan framework for chiplet integration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[11] S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, and S. White, "Pioneering chiplet technology and design for the amd epyc™ and ryzen™ processor families: Industrial product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 57–70.

[12] S. Chen, H. Zhang, Z. Ling, J. Zhai, and B. Yu, "The survey of 2.5d integrated architecture: An eda perspective," in *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, ser. ASPDAC '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 285–293.

[13] A. Coskun, F. Eris, A. Joshi, A. B. Kahng, Y. Ma, and V. Srinivas, "A cross-layer methodology for design and optimization of networks in 2.5 d systems," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.

[14] V. D. Parekh, Z. W. Hazenstab, S. R. Srinivasa, K. Chakrabarty, K. Ni, and V. Narayanan, "Stamp-2.5d: Structural and thermal aware methodology for placement in 2.5d integration," 2025.

[15] W.-H. Liu, M.-S. Chang, and T.-C. Wang, "Floorplanning and signal assignment for silicon interposer-based 3d ics," in *Proceedings of the 51st Annual Design Automation Conference*, 2014, pp. 1–6.

[16] G. Romano, A. Jain, N. Dehmamy, C. Chi, and X. Zhang, "Diffchip: Thermally aware chip placement with automatic differentiation," in *2025 IEEE 75th Electronic Components and Technology Conference (ECTC)*. IEEE, 2025, pp. 221–227.

[17] M. R. Stan, K. Skadron, M. Barcella, W. Huang, K. Sankaranarayanan, and S. Velusamy, "Hotspot: A dynamic compact thermal model at the processor-architecture level," *Microelectronics Journal*, vol. 34, no. 12, pp. 1153–1165, 2003.

[18] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-d multicore systems with stacked dram under power and thermal constraints," in *DAC Design Automation Conference 2012*, 2012, pp. 648–655.

[19] SemiAnalysis. Nvidia's blackwell reworked shipment: Mgx gb200a nvl36 architecture, challenges, and supply chain impact. [Online]. Available: https://semianalysis.com/2024/08/04/nvidias-blackwell-reworked-shipment/

[20] Y. Hsu, M.-H. Chung, Y.-W. Chang, and C.-H. Lin, "Transitive closure graph-based warpage-aware floorplanning for package designs," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–7.

[21] A. Jain, D. P. Parekh, M. Bergendahl, and K. Sikka, "Analytical solution for a chip temperature distribution in a multilayer chip-package with thermal warpage," in *2021 20th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, 2021, pp. 114–122.

[22] D. Sharma *et al.*, "Universal chiplet interconnect express (ucie)," *MEPTEC: Road to Chiplets*, pp. 10–12, 2022.

[23] Z. Yang, S. Ji, X. Chen, J. Zhuang, W. Zhang, D. Jani, and P. Zhou, "Challenges and opportunities to enable large-scale computing via heterogeneous chiplets," *arXiv preprint arXiv:2311.16417*, 2023.

[24] Q. Wang, T. Zhu, Y. Lin, R. Wang, and R. Huang, "Atsim3d: Towards accurate thermal simulator for heterogeneous 3d ic systems considering nonlinear leakage and conductivity," in *2024 International Symposium of Electronics Design Automation (ISEDA)*, 2024, pp. 1–6.

[25] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in vlsi circuits: Principles and methods," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1487–1501, 2006.

[26] B. Wang and P. Mazumder, "Accelerated chip-level thermal analysis using multilayer green's function," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 2, pp. 325–344, 2007.

[27] C. Torregiani, H. Oprins, B. Vandevelde, E. Beyne, and I. De Wolf, "Compact thermal modeling of hot spots in advanced 3d-stacked ics," in *2009 11th Electronics Packaging Technology Conference*, 2009, pp. 131–136.

[28] J. Kung, I. Han, S. Sapatnekar, and Y. Shin, "Thermal signature: A simple yet accurate thermal index for floorplan optimization," in *Proceedings of the 48th Design Automation Conference*, 2011, pp. 108–113.

[29] A. Ziabari, J.-H. Park, E. K. Ardestani, J. Renau, S.-M. Kang, and A. Shakouri, "Power blurring: Fast static and transient thermal analysis method for packaged integrated circuits and power devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 11, pp. 2366–2379, 2014.

[30] S.-Y. Lo, M. Liu, and Y.-W. Chang, "Efficient high-fidelity two-dimensional warpage modeling for advanced packaging analysis," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024, pp. 1–9.

[31] E. Suhir, "Interfacial stresses in bimetal thermostats," *Journal of Applied Mechanics*, vol. 56, no. 3, pp. 595–600, 09 1989.

[32] M.-Y. Tsai and Y.-W. Wang, "A theoretical solution for thermal warpage of flip-chip packages," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 1, pp. 72–78, 2020.

[33] M.-K. Hsu and Y.-W. Chang, "Unified analytical global placement for large-scale mixed-size circuit designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 9, pp. 1366–1378, 2012.

[34] J.-M. Lin, T.-C. Tsai, and R.-T. Shen, "Routability-driven orientation-aware analytical placement for system in package," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 2023, pp. 1–8.

[35] H. Park and O. Chung, "An inverse natural convection problem of estimating the strength of a heat source," *International journal of heat and mass transfer*, vol. 42, no. 23, pp. 4259–4273, 1999.

[36] O. Ciftja, "Electrostatic potential of a uniformly charged square plate at an arbitrary point in space," *Physica Scripta*, vol. 95, no. 9, p. 095802, 2020.

[37] A. Agnesina, P. Rajvanshi, T. Yang, G. Pradipta, A. Jiao, B. Keller, B. Khailany, and H. Ren, "Autodmp: Automated dreamplace-based macro placement," in *Proceedings of the 2023 International Symposium on Physical Design*, 2023, pp. 149–157.

[38] K. Xue, X. Lin, Y. Shi, S. Kai, S. Xu, and C. Qian, "Escaping local optima in global placement," *ArXiv*, vol. abs/2402.18311, 2024.

[39] F. Li, Y. Wang, Y. Cheng, Y. Wang, Y. Han, H. Li, and X. Li, "Gia: A reusable general interposer architecture for agile chiplet integration," in *2022 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2022, pp. 1–9.

[40] J. Nasrullah, Z. Luo, and G. Taylor, "Designing software configurable chips and sips using chiplets and zglue," in *International Symposium on Microelectronics*, vol. 2019, no. 1. International Microelectronics Assembly and Packaging Society, 2019, pp. 000 027–000 032.

[41] T. Zhu, Q. Wang, Y. Lin, R. Wang, and R. Huang, "Efficient warpage simulation of complex 2.5-d/3-d ic structures with novel meshing algorithm and layerwise plate theory," in *2026 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2026.

[42] Q. Wang, T. Zhu, Y. Lin, R. Wang, and R. Huang, "Atsim3.5d: A multiscale thermal simulator for 3.5d-ic systems based on nonlinear multigrid method," in *2025 International Symposium of Electronics Design Automation (ISEDA)*, 2025, pp. 595–601.

[43] Y. Lin, S. Dhar, W. Li, H. Ren, B. Khailany, and D. Z. Pan, "Dreamplace: Deep learning toolkit-enabled gpu acceleration for modern vlsi placement," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[44] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[45] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com

[46] J.-M. Lin, T.-T. Chen, Y.-F. Chang, W.-Y. Chang, Y.-T. Shyu, Y.-J. Chang, and J.-M. Lu, "A fast thermal-aware fixed-outline floorplanning methodology based on analytical models," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.